

Consistency of Neural Causal Partial Identification

Jiyuan Tan ^{*} Jose Blanchet [†] Vasilis Syrgkanis [‡]

October 18, 2026

Abstract

Recent progress in Neural Causal Models (NCMs) showcased how identification and partial identification of causal effects can be automatically carried out via training of neural generative models that respect the constraints encoded in a given causal graph [51, 3]. However, formal consistency of these methods has only been proven for the case of discrete variables or only for linear causal models. In this work, we prove the consistency of partial identification via NCMs in a general setting with both continuous and categorical variables. Further, our results highlight the impact of the design of the underlying neural network architecture in terms of depth and connectivity as well as the importance of applying Lipschitz regularization in the training phase. In particular, we provide a counterexample showing that without Lipschitz regularization this method may not be asymptotically consistent. Our results are enabled by new results on the approximability of Structural Causal Models (SCMs) via neural generative models, together with an analysis of the sample complexity of the resulting architectures and how that translates into an error in the constrained optimization problem that defines the partial identification bounds.

^{*}Management Science and Engineering, Stanford University. Email: jiyuantan@stanford.edu

[†]Management Science and Engineering, Stanford University. Email: jose.blanchet@stanford.edu

[‡]Management Science and Engineering, Stanford University. Email: vsyrgk@stanford.edu. Supported by

NSF Award IIS-2337916 and a 2022 Amazon Research Award

1 Introduction

Identifying causal quantities from observational data is an important problem in causal inference which has wide applications in economics [1], social science [20], health care [34, 18], and recommendation systems [9]. One common approach is to transform causal quantities into statistical quantities using the ID algorithm [48] and deploy general purpose methods to estimate the statistical quantity. However, in the presence of unobserved confounding, typically the causal quantity of interest will not be point-identified by observational data, unless special mechanisms are present in the data generating process (e.g. instruments, unconfounded mediators, proxy controls). In the presence of unobserved confounding, the ID algorithm will typically fail to return a statistical quantity and declare the causal quantity as non-identifiable.

One remedy to this problem, which we focus on in this paper, is partial identification, which aims to give informative bounds for causal quantities based on the available data. At a high level, partial identification bounds can be defined as follows: find the maximum and the minimum value that a target causal quantity can take, among all Structural Causal Models (SCMs) that give rise to the same observed data distribution and respect the given causal graph (as well as any other structural constraints that one is willing to impose). Note that in the presence of unobserved confounding, there will typically exist many structural mechanisms that could give rise to the same observational distribution but have vastly different counterfactual distributions. Hence, partial identification can be formulated as solving a max and a min optimization problem [3]

$$\begin{aligned} & \max_{\mathcal{M} \in \mathcal{C}} \setminus \min_{\mathcal{M} \in \mathcal{C}} \theta(\mathcal{M}), \\ & \text{subject to } P^{\mathcal{M}}(\mathbf{V}) = P^{\mathcal{M}^*}(\mathbf{V}), \quad \text{and} \quad \mathcal{G}_{\mathcal{M}} = \mathcal{G}_{\mathcal{M}^*}, \end{aligned} \tag{1.1}$$

where $\theta(\mathcal{M})$ is the causal quantity of interest, $\mathcal{M}^*$ is the true model, $\mathbf{V}$ is the set of observed nodes, $P^{\mathcal{M}}(\mathbf{V})$ is the distribution of $\mathbf{V}$ in SCM $\mathcal{M}$, $\mathcal{C}$ is a collection of causal models and $\mathcal{G}_{\mathcal{M}}$ is the causal graph of $\mathcal{M}$ (see Section 2 for formal definitions). Two recent lines of work explore the optimization approach to partial identification. The first line deals with discrete Structure Causal Models (SCMs), where all observed variables are finitely supported. In this case, (1.1) becomes a Linear Programming (LP) or polynomial programming problem and tight bounds can be obtained [37, 4, 5, 27, 44, 45, 40, 8, 53, 55, 14]. The second line of work focuses on continuous models and explores ways of solving (1.1) in continuous settings using various techniques [24, 32, 29, 38, 3, 41]. *Remark (notation).* $\mathbf{V} = \{V_i\}_{i=1}^{n_V}$ is the collection of observed variables; joint-law and function-

evaluation notation uses their indexed joint values, and $\mathbf{V}(t)$ is the corresponding collection under intervention $\mathbf{T} = t$. In Proposition C.1: In the proof, $\mathbf{V} = \{X, T, Y\}$ consists of the covariate, treatment, and outcome. $P(X)$ denotes the probability law of X , and $P(A)$ denotes the probability of an event A under the law in context. In Proposition 4.3: Bare P is the reference probability measure on the treatment space used in $P(dt)$; $p_\nu(t|\text{Pa}(T) = x)$ is the density of the conditional treatment law $\nu(dt|x)$ with respect to this same measure. In Corollary 4.4: Bare P is the reference probability measure on the treatment space with respect to which $\nu(dt|\text{Pa}(T) = x) = p_\nu(t|\text{Pa}(T) = x)P(dt)$. The intervention measure satisfies $\mu_T \ll P$ and $\int_{t_1}^{t_2} (d\mu_T/dP(t))^2 P(dt) < \infty$. In a result cited in the proof of Lemma E.6: Bare P is an arbitrary Borel probability distribution on $X \times [0, 1]$, where X is standard Borel. The population squared-loss expectation is taken under this P , and the deviation probability is evaluated in the completion of P^m . Bare $\mathcal{M}$ denotes an individual SCM tuple, whereas $\mathcal{M}(\mathcal{G}, \mathcal{F}, U)$ denotes the class of canonical SCMs with causal graph $\mathcal{G}$, the indicated latent-variable structure, and structural functions $f_i \in \mathcal{F}$ for $i \in [n_V]$. In Theorem 4.2: In the population optimization defining $\underline{F}^{\hat{L}_f}$, $\mathcal{M}(\mathcal{G}, \mathcal{F}_{\hat{L}_f}^K, U)$ uses $\mathcal{F}_L^K = \{f : \sup_x \|f(x)\|_1 \leq K, \text{Lip}(f) \leq L\}$, with nonnegative normalized categorical propensity vectors; the latent-law range consists of independent non-Gumbel laws supported on $[-K, K]^{d_j^U}$ and fixed independent standard Gumbel laws for categorical variables. $\mathcal{G}$ denotes the graph or function class specified by its local occurrence. In Proposition 4.3, Corollary 4.4: $\mathcal{G}$ has no hidden confounder. In Proposition C.1: $\mathcal{G}$ is the common backdoor causal graph of the constructed models, with no unobserved confounding. In Theorem 3.3, Corollary 3.8, Proposition C.1, Proposition 4.1, Theorem 4.2, Proposition 4.3, Corollary 4.4, Proposition B.1, a result cited in the proof of Corollary 3.8: $\mathcal{G}$ is the causal ADMG on $\mathbf{V}$, retaining the structural parent arrows and adding a bidirected arrow between V_i, V_j when correlated latent inputs affect the two nodes. In a result cited in the proof of Lemma E.6: $\mathcal{G}$ is a nonempty class of functions $g : X \rightarrow [0, 1]$ with finite pseudo-dimension v . The covering and packing numbers use evaluations at arbitrary m points of X in the averaged metric $d_1(u, w) = m^{-1} \sum_{i=1}^m |u_i - w_i|$. In a result cited in the proof of Lemma E.6: $\mathcal{G} = \{g(t, \cdot) : t \in T\} \neq \emptyset$, where X, T are standard Borel spaces and $g : T \times X \rightarrow [0, 1]$ is jointly Borel; no pseudo-dimension assumption is imposed. The covering number $\mathcal{N}_1^{\text{avg}}(r, \mathcal{G}, 2m)$ is the supremum over $2m$ points of X of the covering number of the evaluation vectors in the averaged metric $(2m)^{-1} \sum_i |u_i - v_i|$. $\mathcal{C}$ denotes the locally specified collection, consisting of causal models in partial identification and cubes in the geometric construction. In Proposition B.2: $\mathcal{C} = \{C_1, \dots, C_m\}$ is a finite collection of axis-aligned cubes; $\mathcal{C}_n$ is the collection of closed cells in the level- n partition of $[0, 1]^d$, and $\mathcal{C}^* = \bigcup_{n=1}^\infty \mathcal{C}_n$.

θ denotes the causal functional when applied to an SCM in (1.1), and a parameter point in the later optimization and parameterization notation. In Proposition C.2: θ is a point of the metric space (M, d) , ranged over Θ_n , Θ_∞ , or $\tilde{\Theta}_\infty$ in the respective optimization problems. In Theorem 4.2: In the population causal-model parameterization, $\theta = (f_1, \dots, f_{n_V}, P(U_1), \dots, P(U_{n_U})) \in M$, $\mathcal{M}^\theta$ denotes the represented SCM, and P^θ its population distribution; $\theta_{i,j}$ denotes the parameters of the corresponding latent-generator neural network. In the empirical optimization, a candidate index additionally specifies its network representation: different representations with the same latent population laws can give different $P_{m_n}^\theta$ on the common simulation inputs. In Corollary 4.4: In the population causal-model parameterization, $\theta = (f_1, \dots, f_{n_V}, P(U_1), \dots, P(U_{n_U})) \in M$, $\mathcal{M}^\theta$ denotes the represented SCM, and P^θ its population distribution; $\theta_{i,j}$ denotes the parameters of the corresponding latent-generator neural network. In the empirical optimization, a candidate index additionally specifies its network representation: different representations with the same latent population laws can give different $P_{m_n}^\theta$ on the common simulation inputs. $P^\mathcal{M}$ denotes the population-law notation associated with $\mathcal{M}$, evaluated at the variables written in its argument. In Theorem 3.3: $P^\mathcal{M}(\mathbf{V}(\mathbf{t}))$ is the joint interventional law under $\mathbf{T} = \mathbf{t}$, whereas $P^\mathcal{M}(\mathbf{U})$ is the joint latent law. M is the ambient space equipped with the metric d . In Theorem 4.2: In the proof, $M = \mathcal{F}_{V_1} \times \dots \times \mathcal{F}_{V_{n_V}} \times \prod_{j=1}^{n_U} \mathcal{P}([-K, K]^{d_j^U})$. For continuous V_i , $\mathcal{F}_{V_i} = \mathcal{F}_{\hat{L}_f}^K([-K, K]^{d_{i,\text{in}}^V}, [-K, K]^{d_{i,\text{out}}^V})$, with uniform output 1-norm bound K and Lipschitz bound $\hat{L}_f$. For categorical V_i , $\mathcal{F}_{V_i} = \{p : [-K, K]^{d_{i,\text{in}}^V} \rightarrow \Delta_{n_i} : \text{Lip}(p) \leq \hat{L}_f\}$. Here $\mathcal{P}(A)$ is the set of probability measures on A , and $d(\theta, \theta') = \sum_{k=1}^{n_V} \|f_k - f'_k\|_\infty + \sum_{k=1}^{n_U} W(P(U_k), P(U'_k))$.

Recently, Xia et al. [50] formalized the connection between SCMs and generative models (see also [33] for an earlier version of a special case of this connection). This work showcased that SCMs can be interpreted as neural generative models, namely Neural Causal Models (NCMs), that follow a particular architecture that respects the constraints encoded in a given causal graph. Hence, counterfactual quantities of SCMs can be learned by optimizing over the parameters of the underlying generative models. However, there could be multiple models that lead to the same observed data distribution, albeit have different counterfactual distributions. Xia et al. [50] first analyze the approximation power of NCMs for discrete SCMs and employ the max/min approach to verify the identifiability of causal quantities, without the need to employ the ID algorithm. Balazadeh et al. [3] and Hui et al. [30], extend the method in [50] and re-purpose it to perform partial identification by solving the min and max problem in the partial identification formulation over neural causal models.

However, for SCMs with general random variables and functional relationships, the approximation error and consistency of this optimization-based approach to partial identification via NCMs has not been established. In particular, two problems remain open. First, given an arbitrary SCM, it is not yet known if we can find an NCM that produces approximately the same intervention distribution as the original one. Although Xia et al. [50] show it is possible to represent any discrete SCM by an NCM, their construction highly relies on the discrete assumption and cannot be directly generalized to the general case. Moreover, Xia et al. [50] use step functions as the activation function in their construction, which may create difficulties in the training process since step functions are discontinuous with a zero gradient almost everywhere.

Second, since we only have access to n samples from the true distribution, $P^{\mathcal{M}}(\mathbf{V})$, we need to replace the constraints in (1.1) with their empirical version that uses the empirical distribution of samples $P_n^{\mathcal{M}^*}(\mathbf{V})$ in place of the population distribution and looks for NCMs, whose implied distribution lies within a small distance from the empirical distribution. Moreover, even the NCM distribution is typically only accessible through sampling, hence we will need to generate m_n samples from the NCM and use the empirical distribution of the m_n samples from the NCM in place of the true distribution implied by the NCM. Thus, in practice, we will use a constraint of the form $d(P_n^{\mathcal{M}^*}(\mathbf{V}), P_{m_n}^{\mathcal{M}}(\mathbf{V})) \leq \alpha_n$, where d is some notion of distribution distance and α_n accounts for the sampling error. It is not clear that this approach is consistent, converges to the correct partial identification bounds, when the sample size n grows to infinity. Balazadeh et al. [3] only show the consistency of this approach for linear SCMs. Consistency results concerning more general SCMs is still lacking in the neural causal literature.

Remark (notation). n is a local integer parameter whose role is determined by the indexed objects in the surrounding passage. In Lemma B.3, Proposition B.2: In the partition notation $\mathcal{C}_n$ and the curve and refinement sequences, n is a level index beginning at 0, with $\mathcal{C}_0 = \{C_{-1}\}$ and $C_{-1} = [0, 1]^d$. In Lemma E.2: $n \geq 1$ is the number of mixture components, with component indices $k = 1, \dots, n$. In Lemma E.3: $n \geq 1$ is the number of factors in the product measures. In Lemma E.4: $n \geq 1$ is the dimension of X^τ and the number of positive weights p_k and Gumbel coordinates G_k . In Proposition C.2: $n \in \mathbb{N}$ indexes the nested sets, constraint functions, tolerances, and optimization values, and varies in the limit $n \rightarrow \infty$. In Proposition B.2: In the local cube-metric definition $C = \prod_{i=1}^n [a_i, b_i]$ and $C_k = \prod_{i=1}^n [a_i^k, b_i^k]$, n is the ambient dimension and i ranges from 1 to n . In Lemma E.5: $n \geq 1$ is the length of the probability vector $\mu \in \Delta_n$, and a transportation plan has n rows and m columns. In Theorem 4.2, Corollary 4.4: n is the number of observations from the true

distribution and also indexes the approximation and tolerance sequences; the asymptotic limit is $n \rightarrow \infty$. In Lemma E.6: $n \geq 1$ is the sample size defining μ_n and the number of evaluation points in the empirical covering-number notation $\mathcal{N}_1(\epsilon, \mathcal{F}, n)$. In Theorem 3.6: For a component of positive dimension $q = d_i^C$, the refinement integer is $n = \lfloor (W_1/A_q)^{1/q} \rfloor \geq 2$ for $W_1 \geq A_q 2^q$, where $A_q \geq 1$ depends only on q and bounds the clipped network width by $A_q n^q$; the quantization parameter is $\delta = 1/\lceil \sqrt{q}(q+1)n(n-1) \rceil$ and the mesh is δ/n . In a result cited in the proof of Theorem 3.6: For a component of positive dimension $q = d_i^C$, the refinement integer is $n = \lfloor (W_1/A_q)^{1/q} \rfloor \geq 2$ for $W_1 \geq A_q 2^q$, where $A_q \geq 1$ depends only on q and bounds the clipped network width by $A_q n^q$; the quantization parameter is $\delta = 1/\lceil \sqrt{q}(q+1)n(n-1) \rceil$ and the mesh is δ/n . m_n is a prescribed positive integer sequence, indexed by the observational sample size n , giving the simulation sample size for each candidate model. In Theorem 4.2, Corollary 4.4: The simulation sizes satisfy $m_n = \Omega(n)$; equality $m_n = n$ is not required. $\alpha_n = \epsilon_n + s_n$ is a deterministic constraint tolerance sequence indexed by the observational sample size n . In Theorem 4.2: $\alpha_n = \epsilon_n + s_n > 0$, where the positive tolerances ϵ_n and s_n are chosen with sufficiently large constants in their displayed asymptotic orders. In Corollary 4.4: $\alpha_n = \epsilon_n + s_n > 0$, where the positive tolerances ϵ_n and s_n are chosen as sufficiently large constant multiples of the respective quantities in their displayed asymptotic orders. d is local notation for a fixed dimension, a fixed metric, or a regression-difference function, with its role determined by its defining expression. In Proposition B.2: d is the cube dimension in the homeomorphism $f : [0, 1]^d \rightarrow K$ from Assumption 3.5, rather than the ambient dimension D ; the proof treats $d = 0$ separately and uses the exponent $1/d$ only for $d \geq 1$. In Lemma B.3: d is the dimension of the cube $[0, 1]^d$ being partitioned. In Theorem 3.6: d is the ambient dimension of the target law on $\mathbb{R}^d$, distinct from the component cube dimensions d_i^C . In Proposition C.2: $d : M \times M \rightarrow [0, \infty)$ is the metric of the given space (M, d) . In Theorem 4.2: In the proof's parameter space, $d(\theta, \theta') = \sum_{k=1}^{n_V} \|f_k - f'_k\|_\infty + \sum_{k=1}^{n_U} W(P(U_k), P(U'_k))$; occurrences of d in imported distribution bounds retain their local dimension meaning. In Lemma E.6: d is the common output dimension of the functions in $\mathcal{F}$ and the dimension used in the output test-function covering bound; the input law μ is on $\mathbb{R}^{d'}$. In the result split from Theorem 3.6: d is the ambient dimension of the target law on $\mathbb{R}^d$, distinct from the component cube dimensions d_i^C . In Proposition 4.3: In the proof's test-function calculation, $s = (a, t)$ and $d(s) = \widehat{f}_y(s) - f_y(s)$ on the allowed covariate domain Q ; f_y and $\widehat{f}_y$ are the structural regression versions for the respective models. m is a local integer parameter whose role is determined by the objects in the surrounding passage. In Proposition B.2: In $\gamma_m(t) \in C_{k,n-1}$ for all $m \geq n$, m is a later curve level. In

Theorem 4.2: In the Sinkhorn comparison, the second probability-vector length is the simulation count m_n , giving the factor $\log(nm_n)$. In the structural interpolation argument, m is the input dimension of $f : [-K, K]^m \rightarrow \mathbb{R}^{m'}$; the construction treats $m \geq 1$ by triangulation and $m = 0$ by exact constant representation. In Lemma E.5: $m \geq 1$ is the length of $\nu \in \Delta_m$, indexing the points $y_1, \dots, y_m$, and the number of columns of its transportation plan with $\mu \in \Delta_n$; it need not count only positive coordinates or distinct points. In a result cited in the proof of Lemma E.6: $m \geq 1$ is an arbitrary integer counting evaluation points in the averaged empirical covering and packing numbers. In Corollary 4.4: In the structural interpolation argument, m is the input dimension of $f : [-K, K]^m \rightarrow \mathbb{R}^{m'}$; the construction treats $m \geq 1$ by triangulation and $m = 0$ by exact constant representation. In a result cited in the proof of Lemma E.6: $m \geq 1$ is the number of independent observations (x_i, y_i) with law P in the empirical squared loss; $\mathcal{N}_1^{\text{avg}}(\epsilon/16, \mathcal{G}, 2m)$ uses $2m$ evaluation points. In Theorem 3.3: In the recurrence argument, $m = n_V - n_T$ is the number of nonintervened observed nodes. The argument includes $m = 0$, with an empty geometric sum. In Corollary 3.8: In the structural extension and approximation construction, $m = d_{\max}^{\text{in}} > 0$ is the largest structural input dimension, while $m_i = d_i^{\text{in}}$ is the input dimension at node i .

In this paper, we establish representation and consistency results for general SCMs. Our contributions are summarized as follows.

- We show that under suitable regularity assumptions, given any Lipschitz SCM, we can approximate it using an NCM such that the Wasserstein distance between any interventional distribution of the NCM and the original SCM is small. Each random variable of the SCM is allowed to be continuous or categorical. We specify two architectures of the Neural Networks (NNs) that can be trained using common gradient-based optimization algorithms (Theorem 3.6 and Corollary 3.8).
- To construct the NCM approximation, we develop a novel representation theorem of probability measures (Proposition B.2) that may be of independent interest. Proposition B.2 implies that under certain assumptions, probability distributions supported on the unit cube can be simulated by pushing forward a multivariate uniform distribution.
- We discover the importance of Lipschitz regularization by constructing a counterexample where the neural causal approach is not consistent without regularization (Proposition 4.1).
- Using Lipschitz regularization, we prove the consistency of the neural causal approach

(Theorem 4.2).

Related Work There exists a rich literature on partial identification of average treatment effects (ATE) [37, 4, 5, 27, 44, 45, 40, 8, 53, 55, 14, 7, 39, 26]. Balke and Pearl [4, 5] first give an algorithm to calculate bounds on the ATE in the Instrumental Variable (IV) setting. They show that regardless of the exact distribution of the latent variables, it is sufficient to consider discrete latent variables as long as all endogenous variables are finitely supported. Moreover, they discover that (1.1) is an LP problem with a closed-form solution. This LP-based technique was generalized to several special classes of SCMs [44, 8, 27, 46]. For general discrete SCMs, [14, 55, 54] consider transforming the problem (1.1) into a polynomial programming problem. Xia et al. [50] discover the connection between generative models and SCMs. They show that NCMs are expressive enough to approximate discrete SCMs. By setting $\mathcal{C}$ in problem (1.1) to be the collection of NCMs, they apply NCMs for identification and estimation.

For causal models with continuous random variables, the constraints in (1.1) become integral equations, which makes the problem more difficult. One approach is to discretize the constraints. [24] uses stochastic process representation of the causal model in the continuous IV setting and transforms the problem into a semi-infinite program. [32] relaxes the constraints to finite moment equations and solves the problem by the Augmented Lagrangian Method (ALM). The other approach is to use generative models to approximate the observational distribution and use some metric to measure the distance between distributions. [33] first propose to use GAN to generate images. Later, [29] uses Wasserstein distance in the constraint and transforms the optimization problem into a min and max problem. Similarly, [3] solves the optimization problem using Sinkhorn distance to avoid instability during training. They propose to estimate the Average Treatment Derivative (ATD) and use ATD to obtain a bound on the ATE. They also prove the consistency of the proposed estimator for linear SCMs. [41] uses a linear combination of basis functions to approximate response functions. [26] uses sieve to solve the resulting optimization problem in the IV setting. [19] proposes a neural network framework for sensitivity analysis under unobserved confounding.

Organization of this paper In Section 2, we introduce the notations and some basic concepts used throughout the paper. Next, in Section 3, we demonstrate how to construct an NCM so that they can approximate a given SCM arbitrarily well. Two kinds of architecture are given along with an approximation error analysis. In Section 4, we highlight the importance of Lipschitz regularization

by giving a counterexample that is not consistent. Then, leveraging the previous approximation results, we are able to prove the consistency of this approach under regularization. Finally, we compare our method with the traditional polynomial programming method empirically in Section 5.

2 Preliminary

First, we introduce the definition of an SCM. Throughout the paper, we use bold symbols to represent sets of random variables.

Definition 2.1. (Structural Causal Model) A Structural Causal Model (SCM) is a tuple $\mathcal{M} = (\mathbf{V}, \mathbf{U}, F, P(\mathbf{U}), \mathcal{G}_0)$, where $\mathbf{V} = \{V_i\}_{i=1}^{n_V}$ is the set of observed variables; $\mathbf{U} = \{U_j\}_{j=1}^{n_U}$ is the set of latent variables; $P(\mathbf{U})$ is the distribution of latent variables; $\mathcal{G}_0$ is an acyclic directed graph whose nodes are $\mathbf{V}$. The values of each observed variable V_i are generated by

$$V_i = f_i(\text{Pa}(V_i), \mathbf{U}_{V_i}), \quad \text{where } V_i \notin \text{Pa}(V_i) \text{ and } \mathbf{U}_{V_i} \subset \mathbf{U}, \quad (2.1)$$

where $F = (f_1, \dots, f_{n_V})$, $\text{Pa}(V)$ is the set of parents of V in graph $\mathcal{G}_0$ and $\mathbf{U}_{V_i}$ is the set of latent variables that affect V_i . V_i takes either continuous values $\mathbb{R}^{d_i}$ or categorical in $[n_i]$. We extend graph $\mathcal{G}_0$ by adding bi-directed arrows between any $V_i, V_j \in \mathcal{G}_0$ if there exists a correlated latent variable pair $(U_k, U_l), U_k \in \mathbf{U}_{V_i}, U_l \in \mathbf{U}_{V_j}$. We call the extended graph $\mathcal{G}_{\mathcal{M}}$ the causal graph of $\mathcal{M}$.

Remark. For distinct observed vertices V_i, V_j , the “correlated latent variable pair” criterion means that some $U_k \in \mathbf{U}_{V_i}$ and $U_l \in \mathbf{U}_{V_j}$ are not independent, with $k = l$ permitted.

Remark (notation). V_i , for $i \in [n_V]$, denotes the observed variable at node i , and $V_i(\mathbf{t})$ denotes its value under the intervention $\mathbf{T} = \mathbf{t}$, obtained by replacing treatment equations by their assigned values and recursively evaluating the remaining structural equations. In Proposition C.1: In the proof, the observed variables are X, T, Y and the binary treatment T uses the labels 0, 1. For categorical V_i , n_i is its positive category count, with category labels $[n_i] = \{1, \dots, n_i\}$; it is also the length of its propensity vector and G_{V_i} . $U_j, j \in [n_U]$, denotes a member of the indexed latent family; under the categorical convention this family is separate from the categorical noise vectors G_{V_i} . In Theorem 4.2, Corollary 4.4, a result cited in the proof of Theorem 4.2: $U_j, j \in [n_U]$, indexes the non-Gumbel latent coordinates, with $U_j \in [-K, K]^{d_j^U}$ and $D_U = \sum_{j=1}^{n_U} d_j^U$. Their population laws are not required to be uniform; the uniform variables supplied to the latent generators are separate inputs. $\mathbf{U}$ denotes the full latent collection of the SCM, including the separately introduced

categorical generation noises when the categorical convention is used; $U_1, \dots, U_{n_U}$ then index the non-Gumbel part. In Proposition B.1: The original model uses $\mathbf{U} = \{U_1, \dots, U_{n_U}\} \cup \mathbf{G}$, with $\mathbf{G} = \{G_{V_i} : V_i \text{ is categorical}\}$; canonicalization merges the non-Gumbel variables into component blocks and retains the categorical noises. In Theorem 3.3, Corollary 3.8, Theorem 4.2, Proposition 4.3, Corollary 4.4: For canonical SCMs, $\mathbf{U} = \{U_C : C \text{ is a } C^2\text{-component of } \mathcal{G}\} \cup \{G_{V_i} : V_i \text{ is categorical}\}$; the bounded latent coordinates subject to Assumption 3.2 are the non-Gumbel U_j . n_U counts the indexed family $\{U_j\}_{j=1}^{n_U}$, with separately adjoined categorical noises G_{V_i} excluded from that count when the categorical convention is used. In Proposition B.1: n_U counts the original indexed latent variables before merging; the merged latent family is instead indexed by the n_C components. $\mathbf{U}_{V_i}$ is the designated collection of latent inputs to f_i ; under the categorical propensity convention it is a subset of $\{U_1, \dots, U_{n_U}\}$ and excludes the separately displayed categorical Gumbel noise. In Proposition B.1: For the original SCM, $\mathbf{U}_{V_i}$ is its original latent input subset; the constructed canonical SCM uses merged latent blocks, and the reused f_i ignores coordinates outside that original subset. In Theorem 3.3, Corollary 3.8, Theorem 4.2, Proposition 4.3, Corollary 4.4: For the canonical SCMs in $\mathcal{M}(\mathcal{G}, \mathcal{F}, \mathbf{U})$, $\mathbf{U}_{V_i} = \{U_C : V_i \in C, C \text{ is a } C^2\text{-component of } \mathcal{G}\}$. $\text{Pa}(V_i)$ is the set of directed parents of V_i in $\mathcal{G}_0$, excludes V_i , and in structural-function arguments or conditioning expressions denotes the corresponding collection of parent values. f_i , for $i \in [n_V]$, is the function associated with node V_i , evaluated on the values of its observed parents and designated latent inputs; under the categorical convention it denotes the nonnegative propensity vector rather than the categorical label. In Corollary 3.8, a result cited in the proof of Corollary 3.8: The neural approximation argument uses the stated cube-domain convention for the inputs of f_i , with d_i^{in} and d_i^{out} its input and output dimensions. In Proposition B.1: When the proof reuses f_i after merging latent variables into blocks, it denotes the original function applied to the original coordinates and ignores added block coordinates that were not in $\mathbf{U}_{V_i}$. For categorical nodes the original weights satisfy $f_i(x) \in [0, \infty)^{n_i}$ and $\|f_i(x)\|_1 > 0$ on their structural input domain, with $\log 0 = -\infty$; the constructed canonical categorical function is the normalized vector $f_i/\|f_i\|_1$ with coordinates regrouped. Its smoothness is that of this normalized vector: normalization preserves differentiability order and local Lipschitz regularity wherever the total weight is positive, and regrouping preserves any global Lipschitz regularity the normalized vector possesses. In Theorem 3.3: In $\|f_i - \hat{f}_i\|_\infty$, the structural uniform infinity norm is the supremum of the coordinatewise infinity norm over the common structural input domain; zero categorical coordinates use $\log 0 = -\infty$ in the canonical argmax construction. In Theorem 4.2: The proof

treats f_i as an element of $\mathcal{F}_{V_i}$ on $[-K, K]^{d_{i,\text{in}}^V}$ with bounded output cube $[-K, K]^{d_{i,\text{out}}^V}$, imposing $\|f_i(x)\|_1 = 1$ pointwise for categorical nodes; the population bounded Lipschitz class is read as $\mathcal{F}_L^K = \{f : \sup_x \|f(x)\|_1 \leq K, \text{Lip}(f) \leq L\}$. In Theorem 3.3, Corollary 3.8, Theorem 4.2, Proposition 4.3, Corollary 4.4, a result cited in the proof of Corollary 3.8: For the canonical SCM functions in (3.1), categorical f_i are normalized: $(f_i)_k \geq 0$ and $\|f_i\|_1 = 1$, so their values lie in Δ_{n_i} and their coordinates are the conditional category probabilities given $\text{Pa}(V_i), \mathcal{U}_{V_i}$. In Theorem 4.2, Corollary 4.4: For every continuous structural function f_i of $\mathcal{M}^*$, $\sup_x \|f_i(x)\|_1 \leq K$ on its full structural input domain, and every continuous candidate structural network $\hat{f}_i \in \mathcal{F}_{0,n}$ satisfies the same uniform output 1-norm bound. F is functional notation whose tuple, scalar-objective, and push-forward-map meanings are distinguished by context. In Theorem 3.3, Corollary 3.8, Proposition C.1, Proposition 4.1, Proposition 4.3: In the SCM notation, $F = (f_1, \dots, f_{n_V})$ is the tuple of functions in the structural equations. In Theorem 4.2: In (4.1) and (4.2), $F : \mathbb{R}^{n_V} \rightarrow \mathbb{R}$ is the Lipschitz continuous causal objective evaluated on the intervened observed variables, rather than the tuple of structural functions. In Corollary 4.4: The causal objective is the outcome coordinate projection $F(\mathbf{V}) = Y$. In Lemma E.1: $F : K \rightarrow \mathbb{R}^{d_2}$ is a measurable, L -Lipschitz continuous map, where $K \subset \mathbb{R}^{d_1}$ is compact. In Proposition B.1: The SCM tuple uses $F = (f_1, \dots, f_{n_V})$; in the proof's expression $\mathbf{V} = F(\mathbf{U})$, F instead denotes the map obtained by recursively evaluating the structural equations, with its intervention version denoted F_t . f denotes a locally specified function, with domain, codomain and regularity supplied by its occurrence, rather than one common map shared across the paper. In Proposition C.2: $f : \Theta_\infty \rightarrow \mathbb{R}$ is the continuous objective minimized in all three optimization problems. In Lemma C.3: $f(t) = \mathbb{E}_\nu[Y \mid T = t]$ is the population regression function, taken in the L -Lipschitz version specified by the hypothesis, with $|f(t)| \leq K$. In Lemma E.2: In the proof, $f : \mathbb{R}^d \rightarrow \mathbb{R}$ ranges over 1-Lipschitz test functions in the Wasserstein dual formulation. In Lemma E.5: In the proof, $f(x) = -x \log x$ is the concave scalar function used on transportation-plan entries in $[0, 1]$, with its continuous value $f(0) = 0$. In Theorem 3.3: In the categorical step of the proof, the unindexed f denotes the structural propensity function at the node currently being compared, and $p = f(\text{pa}(v_k), \mathbf{u}_{V_k})$. In Lemma E.6: f ranges over the fixed class $\mathcal{F}$ of measurable maps $\mathbb{R}^{d'} \rightarrow \mathbb{R}^d$, with $\|f(x)\|_\infty \leq K$ for every x ; each scalar coordinate class $\mathcal{F}_j = \{x \mapsto f_j(x) : f \in \mathcal{F}\}$, $j \in [d]$, has pseudo-dimension less than $d_{\mathcal{F}} < \infty$. There are a standard Borel space T and a jointly Borel map $G : T \times \mathbb{R}^{d'} \rightarrow [-K, K]^d$ such that $\mathcal{F} = \{G(t, \cdot) : t \in T\}$. Probabilities of uniform deviation events are evaluated in the completion of the sample law. In the result split from Theorem 3.6: In the lower-bound assumption,

$f \in \mathcal{H}([0, 1]^d, K) \cap \mathcal{F}_L([0, 1]^d, K)$ is a Lipschitz homeomorphism satisfying $\mathbb{P}(f(B)) \geq C_f \lambda(B)$ for measurable $B \subset [0, 1]^d$; when applied componentwise, the support and measure are those of the component. In Proposition B.2, Corollary 3.8, Theorem 4.2, Corollary 4.4: In the lower-bound assumption, $f \in \mathcal{H}([0, 1]^d, K) \cap \mathcal{F}_L([0, 1]^d, K)$ is a Lipschitz homeomorphism satisfying $\mathbb{P}(f(B)) \geq C_f \lambda(B)$ for measurable $B \subset [0, 1]^d$; when applied componentwise, the support and measure are those of the component. In Theorem 4.2: In the proof's optimization parameterization, $f(\theta) = \mathbb{E}_{t \sim \mu_T} \mathbb{E}_{\mathcal{M}^\theta} [F(V_1(t), \dots, V_{n_V}(t))]$ is the population intervention objective on the model parameter space, distinct from the structural functions f_i and the homeomorphism in the lower-bound assumption. In Proposition 4.1, Proposition C.1: For each non-Gumbel latent law $\mathbb{P}$ and each connected support component C_i of positive mass, the lower-bound assumption uses a Lipschitz homeomorphism $f \in \mathcal{H}([0, 1]^d, C_i) \cap \mathcal{F}_L([0, 1]^d, C_i)$ satisfying $\mathbb{P}(f(B))/\mathbb{P}(C_i) \geq C_f \lambda(B)$ for every measurable $B \subset [0, 1]^d$.

When we write $\text{Pa}(V_i)$, we refer to the parents of V_i in the $\mathcal{G}_0$. To connect with the literature, the causal graph we define is a kind of Acyclic Directed Mixed Graph (ADMG), which is often used to represent SCMs with unobserved variables [48]. Note that we allow one latent variable to enter several nodes, which differs from the common definition. We use n_U and n_V to denote the number of latent variables and observable variables. Let $\mathbf{T} \subset \mathbf{V}$ be a set of treatment variables. The goal is to estimate causal quantities under a given intervention $\mathbf{T} = \mathbf{t}$. Formally, the structural equations of the intervened model are

$$\begin{aligned} T_i &= t_i, \quad \forall T_i \in \mathbf{T}, \\ V_i(t) &= f_i(\text{Pa}(V_i), \mathbf{U}_{V_i}), \quad \forall V_i \notin \mathbf{T}. \end{aligned}$$

We denote $V_i(\mathbf{t})$ to be the value of V_i under the intervention $\mathbf{T} = \mathbf{t}$ and $P^\mathcal{M}(\mathbf{V}(\mathbf{t}))$ to be the distribution of $\mathbf{V}(\mathbf{t})$ in $\mathcal{M}$. The notion of a C^2 component [50] is defined as follows.

Definition 2.2 (C^2 -Component). For a causal graph $\mathcal{G}$, a subset $C \subset \mathbf{V}$ is C^2 -component if each pair $V_i, V_j \in C$ is connected by a bi-directed arrow in $\mathcal{G}$ and C is maximal.

Remark. C is a C^2 -component precisely when it is inclusion maximal among subsets of $\mathbf{V}$ whose distinct vertices are pairwise joined by a bidirected arrow.

We provide a concrete example in Section A to explain all these notions. We will make the following standard assumption about the independence of latent variables. Note that since we allow latent variables to enter in multiple structural equations, this is more a notational convention and not

an actual assumption. Also note that under this convention a bi-directed arrow essentially represents the existence of a common latent parental variable.

Remark (notation). The meaning of C is determined by its local definition: a graph component, a geometric set, a numerical constant, or the complete categorical label vector. In Theorem 3.3, Corollary 3.8, Proposition C.1, Proposition 4.1, Theorem 4.2, Proposition 4.3, Corollary 4.4, Proposition B.1: In canonical latent indexing, $C \subset V$ is maximal subject to every pair of distinct vertices in C being joined by a bi-directed arrow in $\mathcal{G}$. In Theorem 3.6, Corollary 3.8, Theorem 4.2, Corollary 4.4: In the distribution-generator construction, C_i denotes a connected component of $\text{supp}(\mathbb{P})$, with $\text{supp}(\mathbb{P}) = \bigcup_{i=1}^{N_C} C_i$ and $\mathcal{H}([0, 1]^{d_i}, C_i) \cap \mathcal{F}_L([0, 1]^{d_i}, C_i) \neq \emptyset$. In Proposition B.2: In the partition construction, $C = \prod_{i=1}^n [a_i, b_i]$ is an axis-aligned cube; indexed sets $C_{i_1, \dots, i_n}$ are the closed cells of the constructed partitions of $[0, 1]^d$. In Lemma E.6: $C > 0$ is a deterministic numerical constant depending only on d, K in the probability bound. In Theorem 4.2, Corollary 4.4: In numerical probability and tolerance bounds, $C > 0$ denotes a deterministic constant rather than a set of observed vertices. In Proposition 4.1, Proposition C.1: In the lower-bound premise for each non-Gumbel latent law $\mathbb{P}$, C_i denotes a connected component of its support, and the assumption is applied with $K = C_i$ to the normalized component law $\mathbb{P}|_{C_i}/\mathbb{P}(C_i)$ wherever this normalization is defined. In Theorem 4.2, a result cited in the proof of Theorem 4.2: In the latent empirical concentration argument, C denotes the complete categorical label vector, with values $c \in \prod_{i: V_i \text{ is categorical}} [n_i]$; $p_c = P(C = c)$, N_c is its observational sample count, and $\mu_c = P((U_1, \dots, U_{n_U}) \mid C = c)$ for $p_c > 0$. $\mathbf{t} = (t_i)_{T_i \in T}$ assigns treatment values to the chosen set $T \subset V$ by replacing each treatment structural equation with $T_i = t_i$. In Proposition C.1, Proposition 4.1: For the displayed ATE contrast, $T = \{T\}$ and the assigned treatment value is 1. In Proposition 4.3: The intervention is on the single treatment T , and its value t ranges over $[t_1, t_2]$ in the displayed integral. In Corollary 3.8: For the fixed $R > 0$ chosen before $\widehat{\mathcal{M}}$, each continuous treatment coordinate satisfies $\|t_i\|_\infty \leq R$, and each categorical treatment coordinate belongs to its declared set $[n_i]$. T is locally specified as a treatment variable, a transportation map, a transportation plan, or a standard Borel parameter space. In Lemma C.3: T is the treatment coordinate of $(T, Y) \sim \nu$; conditioning in $f(t)$ is on $T = t$. In Proposition C.1: T is the binary treatment in the proof, taking values 0 and 1, with the conditional probabilities specified by the displayed structural equations. In Lemma E.4: $T : \Delta_n \rightarrow \Delta_n$ sends x to the one-hot vector at the first coordinate attaining $\max_j x_j$; it is a transportation map, rather than a treatment variable. In Proposition 4.3, Corollary 4.4: T denotes the treatment node, and $Y(t)$ denotes the outcome

under the fixed intervention $T = t$. In Lemma E.5: $T = (T_{ij}) \in \Pi(\mu, \nu)$ is a nonnegative $n \times m$ transportation plan satisfying $\sum_j T_{ij} = \mu_i$ and $\sum_i T_{ij} = \nu_j$. The minimizations defining the transport costs range over $\Pi(\mu, \nu)$; the entropy maximum displayed in the proof ranges over plans attaining the unregularized optimal cost $W(\mu, \nu)$. In Lemma E.6: T is a standard Borel parameter space, and $G : T \times \mathbb{R}^{d'} \rightarrow [-K, K]^d$ is jointly Borel with $\mathcal{F} = \{G(t, \cdot) : t \in T\}$. It is not a treatment variable. In a result cited in the proof of Lemma E.6: In the cited squared-loss formulation, T is a standard Borel parameter space for a jointly Borel map $g : T \times X \rightarrow [0, 1]$, with $\mathcal{G} = \{g(t, \cdot) : t \in T\} \neq \emptyset$.

Assumption 2.3. All the latent variables in U are independent.

To deal with categorical variables, we assume that latent variables that influence categorical variables contain two parts: the shared confounding that influences the propensity functions and the independent noise that generates categorical distributions.

Assumption 2.4. The set of latent variables consists of two parts $U = \{U_1, \dots, U_{n_U}\} \cup \{G_{V_i} : V_i \text{ is categorical}\}$. Precisely, if $V_i \in \mathbf{V}$ is a categorical variable, the data generation process of V_i satisfies $V_i = \arg \max_{k \in [n_i]} \{g_k^{V_i} + \log(f_i(\text{Pa}(V_i), U_{V_i}))_k\} \sim \text{Categorical}(f_i / \|f_i\|_1)$, where $G_{V_i} = (g_1^{V_i}, \dots, g_{n_i}^{V_i})$ are i.i.d. standard Gumbel variables, $U_{V_i} \subset \{U_1, \dots, U_{n_U}\}$.

Remark. Writing $x = (\text{Pa}(V_i), U_{V_i})$, the displayed law means $P(V_i = k \mid x) = (f_i(x))_k / \|f_i(x)\|_1$ for the nonnegative weight vector defining the categorical distribution, with zero coordinates interpreted using $\log 0 = -\infty$.

Remark (notation). $G_{V_i} = (g_1^{V_i}, \dots, g_{n_i}^{V_i})$ is present for categorical V_i and consists of standard Gumbel coordinates furnishing that node's categorical-generation noise, separately from the latent inputs U_{V_i} of its propensity function. g denotes a real-valued function on the domain specified by its local use. In Theorem 3.3: In the proof, g ranges over real-valued 1-Lipschitz Wasserstein dual test functions of $(\mathbf{u}, \mathbf{t}, v_1, \dots, v_k)$. In Lemma E.1: In the proof, $g : \mathbb{R}^{d_2} \rightarrow \mathbb{R}$ is an arbitrary 1-Lipschitz test function. In Lemma C.3: In the proof, g ranges over real-valued 1-Lipschitz test functions of (t, y) in the Wasserstein dual formula. In Proposition C.2: $g : \Theta_\infty \rightarrow \mathbb{R}$ is the limiting constraint, with $\sup_{\theta \in \Theta_\infty} |g_n(\theta) - g(\theta)| \leq \delta_n$; its zero set defines the limiting feasible regions. In Theorem 4.2: $g(\theta) = W(P^{\theta^*}(\mathbf{V}), P^\theta(\mathbf{V}))$ is the population observational discrepancy, rather than an empirical discrepancy. In a result cited in the proof of Lemma E.6: g ranges over functions in a nonempty class $\mathcal{G} \subset \{g : X \rightarrow [0, 1]\}$ with finite pseudo-dimension v . No Lipschitz condition on g

is specified. In a result cited in the proof of Lemma E.6: $g : T \times X \rightarrow [0, 1]$ is jointly Borel, where T and X are standard Borel spaces, and $\mathcal{G} = \{g(t, \cdot) : t \in T\} \neq \emptyset$. For each $t \in T$, the squared loss is $(g(t, x) - y)^2$. No pseudo-dimension bound or Lipschitz condition on g is imposed.

This convention is without loss of generality at this point, but will be useful when introducing Lipschitz restrictions on the structural equation functions. The Gumbel variables in the assumption can be replaced by any random variables that can generate categorical variables. It can be proven that all discrete SCMs satisfy this assumption. Note that we implicitly assume that all categorical variables V_i are supported on $[n_i]$ for some n_i . It is straightforward to generalize all results to any finite support. Next, we introduce Neural Causal Models (NCMs).

Definition 2.5. (Neural Causal Model) A Neural Causal Model (NCM) is a special kind of SCM where $\mathbf{U} = \{U_1, \dots, U_{n_U}\} \cup \{G_{V_i} : V_i \text{ categorical}\}$, the scalar coordinates of all the U_i are mutually independent with common $\text{Unif}([0, 1])$ law, and U_i groups the indicated finite number d_i^U of these coordinates, so that $U_i \sim \text{Unif}([0, 1]^{d_i^U})$. All coordinates of the G_{V_i} are independent standard Gumbel variables, independent of all uniform coordinates. The continuous structural maps and the categorical propensity-weight maps f_i in (2.1) are Neural Networks (NNs); categorical values are generated from these weights by the Gumbel argmax rule. The weights are nonnegative and have strictly positive sum, with $\log 0 = -\infty$.

Remark (notation). U_i denotes the index- i random variable in the locally specified latent or generator-input family. In Corollary 3.8: In the SCM notation, U_i is a non-Gumbel latent variable of the model being approximated, with its own latent dimension and population law; scalar uniform variables appearing in the component-generator construction are inputs used to approximate that law. In Theorem 3.6, the result split from Theorem 3.6: In the component-generator construction, $U_i \sim \text{Unif}([0, 1])$ for $i \in [N_C]$; i indexes connected support components, and U_i is the scalar input to $\hat{g}_1^i$. In a result cited in the proof of Theorem 3.6: In the component-generator construction, $U_i \sim \text{Unif}([0, 1])$ for $i \in [N_C]$; i indexes connected support components, and U_i is the scalar input to $\hat{g}_1^i$. In a result cited in the proof of Theorem 4.2: $U_i, i \in [n_U]$, is a non-Gumbel latent variable with values in $[-K, K]^{d_i^U}$ under the locally specified model law P^θ or P^{θ^*} ; the scalar uniform generator inputs are $Z_{i,j}$. The vector $(U_1, \dots, U_{n_U})$ has total dimension $D_U = \sum_{i=1}^{n_U} d_i^U$.

The definition of NCMs we use is slightly different from that in [51] because we need to deal with mixed variables in our models. In (1.1), we usually take $\mathcal{C}$ to be the set of all SCMs. However, it is difficult to search over all SCMs since (1.1) becomes an infinite-dimensional polynomial

programming problem. As an alternative, we can search over all NCMs. One quantity of common interest in causal inference is the Average Treatment Effect (ATE).

Definition 2.6. (Average Treatment Effect). For SCM $\mathcal{M}$, the ATE at $\mathbf{T} = \mathbf{t}$ with respect to $\mathbf{T} = \mathbf{t}_0$ is given by $\text{ATE}_{\mathcal{M}}(\mathbf{t}) = \mathbb{E}_{\mathbf{u} \sim P(\mathbf{U})}[Y(\mathbf{t}) - Y(\mathbf{t}_0)]$.

Remark (notation). $\mathbf{t}_0$ assigns the reference values to the same treatment set $\mathbf{T}$ as $\mathbf{t}$ in $\mathbb{E}_{\mathbf{u} \sim P(\mathbf{U})}[Y(\mathbf{t}) - Y(\mathbf{t}_0)]$. In Proposition C.1, Proposition 4.1: The reference intervention is $T = 0$. Y denotes the random quantity specified by the local model or joint law, with the outcome and coupling uses distinguished below. In Proposition C.1, Proposition 4.1, Proposition 4.3, Corollary 4.4: Y is the scalar outcome node of the SCM, and $Y(t)$ is its value after replacing the treatment equation by $T = t$. In Lemma C.3: Y is the second coordinate of $(T, Y) \sim \nu$, and $f(t)$ is a version of $\mathbb{E}_{\nu}[Y \mid T = t]$. In Corollary 4.4: The objective selects the outcome coordinate, $F(\mathbf{V}) = Y$. In Lemma E.3: In the proof, $Y = (Y_1, \dots, Y_n)$ is the second product-coupled vector; $(X_k, Y_k) \sim \pi_k$, where π_k couples μ_k and $\hat{\mu}_k$, so Y has law $\hat{\mu}$. $\text{ATE}_{\mathcal{M}}(\mathbf{t}) = \mathbb{E}_{\mathbf{u} \sim P(\mathbf{U})}[Y(\mathbf{t}) - Y(\mathbf{t}_0)]$ is a population contrast between the two intervened outcomes in the same SCM. In Proposition C.1, Proposition 4.1: The argument suppressed in $\text{ATE}_{\mathcal{M}}$ is $T = 1$, and the reference intervention is $T = 0$.

Partial identification can be formulated as estimating the solution to the optimization problems (1.1) [3]. The max and min values $\overline{F}$ and $\underline{F}$ define the interval $[\underline{F}, \overline{F}]$ which is the smallest interval we can derive from the observed data without additional assumptions. In particular, if $\overline{F} = \underline{F}$, then the causal quantity is point-identified.

Notations. We use $\|\cdot\|$, $\|\cdot\|_{\infty}$ for the 1-norm and ∞ -norm and $[n]$ for $\{1, \dots, n\}$. Bold letters represent sets of random variables. We let $\mathcal{F}_L(K_1, K_2)$ be the class of Lipschitz L -continuous functions $f : K_1 \rightarrow K_2$. We may omit the domain and use $\mathcal{F}_L$ when the domain is clear from context. Let $\mathcal{H}(K_1, K_2)$ be the set of homeomorphisms from K_1 to K_2 , i.e., injective and continuous maps in both directions. We define $\epsilon(\mathcal{F}_1, \mathcal{F}_2) = \sup_{f_2 \in \mathcal{F}_2} \inf_{f_1 \in \mathcal{F}_1} \|f_2 - f_1\|$ for function classes $\mathcal{F}_1, \mathcal{F}_2$. We use standard asymptotic notation $O(\cdot), \Omega(\cdot)$. Given a measure μ on $\mathbb{R}^{d_1}$ and a measurable function $f : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$, the push-forward measure $f_{\#}\mu$ is defined as $f_{\#}\mu(B) = \mu(f^{-1}(B))$ for all measurable sets B . We use $P(X)$ to represent the distribution of random variable X . Let $\Delta_n = \{(p_1, \dots, p_n) : \sum_{i=1}^n p_i = 1, p_i \geq 0\}$ be the probability simplex. We use $\text{Categorical}(\mathbf{p}), \mathbf{p} \in \Delta_n$ to represent categorical distribution with event probability $\mathbf{p}$. We let $W(\cdot, \cdot)$ be the Wasserstein-1 distance and $S_{\lambda}(\mu, \nu)$ be the Sinkhorn distance [12] with regularization parameter $\lambda > 0$.

Remark (notation). $\epsilon(\mathcal{F}_1, \mathcal{F}_2)$ denotes the directed approximation error $\sup_{f_2 \in \mathcal{F}_2} \inf_{f_1 \in \mathcal{F}_1} \|f_2 - f_1\|$,

whereas a bare $\epsilon > 0$ denotes a local positive real tolerance. In Proposition 4.1, Proposition C.1: ϵ is arbitrary and positive, and specifies the requested upper bound on $W(P^{\mathcal{M}^*}(\mathbf{V}), P^{\mathcal{M}_\epsilon}(\mathbf{V}))$. In Lemma E.3: In the proof, ϵ is the positive slack in each chosen coupling's transportation cost and is subsequently taken to zero. In Proposition C.2: In the proof, ϵ is a positive gap witnessing a hypothetical subsequence with $f_{n_k}^* < f^* - \epsilon$. In Corollary 4.4: The statement introduces $\epsilon > 0$ without assigning it any further role. In the empirical concentration argument of the proof, a local deviation tolerance is chosen as $\epsilon = Cr_n$, together with $\delta = r_n$, for sufficiently large n . In Lemma E.6: $\epsilon > 0$ is the deviation tolerance and also the unnormalized covering radius when it appears in $\mathcal{N}_1(\epsilon, \mathcal{F}, n)$; the probability bound is for each fixed pair $\delta, \epsilon > 0$. In the proof, for $0 < \epsilon < 2dK$, the normalized squared-loss tolerance is $\eta = \epsilon/(2dK)$. In a result cited in the proof of Lemma E.6: $0 < \epsilon \leq 1$ is the covering and packing radius in the averaged metric $d_1(u, w) = m^{-1} \sum_{i=1}^m |u_i - w_i|$. In a result cited in the proof of Lemma E.6: $0 < \epsilon < 1$ is the deviation tolerance for the absolute difference between population and empirical squared loss; the covering radius is $\epsilon/16$ on $2m$ evaluation points. In Corollary 3.8: In the scalar inequality $|\max\{\epsilon, z\} - p| \leq \max\{\epsilon, |z - p|\}$, $\epsilon \geq 0$ is an arbitrary floor parameter; the inequality is then applied coordinatewise with $\epsilon = \epsilon_i$. In Theorem 4.2: In the empirical concentration argument, the local deviation tolerance is chosen as $\epsilon = A_0 r_n$, together with $\delta = r_n$, for sufficiently large n . μ denotes a measure on the domain specified by its local use, rather than a single source measure on a globally fixed Euclidean space. In Lemma E.1: μ is a probability measure supported on the compact set $K \subset \mathbb{R}^{d_1}$. In Lemma E.3: $\mu = \mu_1 \otimes \cdots \otimes \mu_n$ is the product of the stated component measures. In Lemma C.3: μ is the joint law of $(\widehat{T}, \widehat{Y})$, and $\widehat{f}(t)$ is the conditional expectation under this law. In Proposition 4.3: μ is the observational distribution of $\widehat{\mathcal{M}}$. In Lemma E.5: $\mu \in \Delta_n$ is the weight vector of a finite probability measure on the supplied points $x_1, \dots, x_n$ in the common Euclidean ground space. In Lemma E.6: μ is the population probability measure on $\mathbb{R}^{d'}$ whose empirical distribution is μ_n . ν denotes a probability distribution, represented by its probability vector when its domain is finite, with the ambient domain specified locally. In Proposition 4.3: $\nu = P^{\mathcal{M}}(\mathbf{V})$ is the observational law of $\mathcal{M}$. In the proof, ν_A is its marginal on the observed nondescendants A of T , excluding T , and $X = \text{Pa}(T)$ is a subvector of A . The conditional treatment laws satisfy $\nu(dt \mid A = a) = \nu(dt \mid X = x) = \nu_\nu(t \mid X = x)P(dt)$, where x is the X subvector of a . In Lemma C.3: $\nu = P(T, Y)$ is the joint law of (T, Y) , while $\nu(dt)$ in the integral denotes its treatment marginal. In Lemma E.5: $\nu = (\nu_1, \dots, \nu_m) \in \Delta_m$ is the weight vector representing the finite probability measure $\sum_{j=1}^m \nu_j \delta_{y_j}$; the points y_j and the points x_i representing μ lie in a common

Euclidean space, with ground distance $\|x_i - y_j\|$ given by the 1-norm. In Theorem 3.6, the result split from Theorem 3.6: ν^τ is the distribution of the Gumbel-softmax weight vector w^τ , and ν^0 is the distribution of its limiting one-hot vector. The notation $W(\mu, \nu)$ denotes Wasserstein-1 distance between measures on the same underlying space, whereas W in a network architecture argument denotes an integer width. In a result cited in the proof of Corollary 3.8: In $a\omega(cW^{-2/d_i^{\text{in}}})$, W is the number of weights of the scalar approximating network, rather than its width; at fixed width $2d_i^{\text{in}} + 10$, this weight count grows linearly in the depth L_0 . λ is a positive scalar when it indexes a regularized transport functional and a reference measure when it occurs in a push-forward or measure-domination expression, with its reference domain specified locally. In Theorem 3.6: In the push-forward expression, λ is Lebesgue measure on $[0, 1]^{N_C}$; in the componentwise construction, it is the uniform measure on $[0, 1]$. In Proposition B.2: For the curve’s push-forward, λ is Lebesgue measure on $[0, 1]$; in expressions $\lambda(B)$ with $B \subset [0, 1]^d$, it is d -dimensional Lebesgue measure. In Lemma E.5: $\lambda > 0$ is the scalar regularization parameter. In the result split from Theorem 3.6: In the push-forward expression, λ is Lebesgue measure on $[0, 1]^{N_C}$; in the componentwise construction, it is the uniform measure on $[0, 1]$, and in the lower-bound premise it is Lebesgue measure on the corresponding cube. S_λ is the debiased regularized transport functional obtained by subtracting half of each self-comparison from $W_{1,\lambda}(\mu, \nu)$. In Lemma E.5: The arguments are finite probability measures represented by their weights and support points, with $c_{ij} = \|x_i - y_j\|$ for the 1-norm. Here $W_{1,\lambda}(\mu, \nu) = \min_{T \in \Pi(\mu, \nu)} \{ \sum_{i,j} c_{ij} T_{ij} - \lambda h(T) \}$, where $\Pi(\mu, \nu)$ is the set of nonnegative transportation plans with marginals μ, ν , and $h(T) = 1 - \sum_{i,j} T_{ij} \log T_{ij}$ with $0 \log 0 = 0$. Both self terms use this same convention and the same 1-norm ground distance on their corresponding support points. L denotes a nonnegative Lipschitz bound with respect to the paper’s 1-norm when used in a Lipschitz condition or $\mathcal{F}_L$, and an integer depth when used as the depth argument of a neural network class. In Theorem 3.3, Corollary 3.8, Proposition 4.3: In the SCM class $\mathcal{M}(\mathcal{G}, \mathcal{F}_L, \mathcal{U})$, L bounds the continuous structural functions and the categorical propensity functions, rather than the categorical arg max maps. In Lemma E.1: L is a Lipschitz bound for F ; the hypothesis does not impose that bound on $\hat{F}$. In Lemma C.3: L is a common Lipschitz bound for the regression functions f and $\hat{f}$. In Proposition B.2, Theorem 3.6: In the support assumptions, L bounds the homeomorphisms from the unit cubes to the support sets; it is not a Lipschitz bound for the resulting space-filling curves. In Theorem 4.2: In $\mathcal{F}_L^K$ and $\underline{F}^L$, L bounds the Lipschitz constants of the continuous structural functions and normalized categorical propensity vectors of population candidate SCMs, with respect to the paper’s 1-norm; the class also requires $\sup_x \|f(x)\|_1 \leq K$ on the full structural

input domain. The conclusion evaluates this parameter at $\widehat{L}_f$. In the result split from Theorem 3.6: In the support assumptions, L bounds the homeomorphisms from the unit cubes to the support sets; it is not a Lipschitz bound for the resulting space-filling curves. $\mathcal{F}_L(K_1, K_2)$ denotes the class of maps $f : K_1 \rightarrow K_2$ with Lipschitz constant at most L , using the paper’s 1-norm convention, with the arguments supplied by context when omitted. $\mathcal{H}(K_1, K_2)$ denotes homeomorphisms from K_1 onto K_2 , while an explicit local definition of unparameterized $\mathcal{H}$ supplies its function class. In Lemma E.6: In the proof, $\mathcal{H} = \{h : h \in \mathcal{F}_1([-K, K]^d, [-K, K]^d), h(0) = 0\}$, and $\mathcal{H}_\delta$ is a δ -net of this class in the ℓ_∞ norm. $\mathcal{F}$ denotes a function class whose domains and codomains are supplied by its local use. In Theorem 3.3, Corollary 3.8, Proposition C.1, Proposition 4.1, Theorem 4.2, Proposition 4.3, Proposition B.1: In $\mathcal{M}(\mathcal{G}, \mathcal{F}, U)$, $\mathcal{F}$ specifies the allowed structural functions, with $f_i \in \mathcal{F}$ for $i \in [n_V]$; each function has the input and output dimensions of its structural equation. In Lemma E.6: $\mathcal{F}$ is a class of measurable maps $f : \mathbb{R}^{d'} \rightarrow \mathbb{R}^d$. Each scalar coordinate class $\mathcal{F}_j = \{x \mapsto f_j(x) : f \in \mathcal{F}\}$, $j \in [d]$, has pseudo-dimension less than $d_{\mathcal{F}} < \infty$, and $\|f(x)\|_\infty \leq K$ for every $f \in \mathcal{F}$ and $x \in \mathbb{R}^{d'}$. There are a standard Borel space T and a jointly Borel map $G : T \times \mathbb{R}^{d'} \rightarrow [-K, K]^d$ such that $\mathcal{F} = \{G(t, \cdot) : t \in T\}$. K is a context-dependent bound or compact set: numerical occurrences are positive bounds, whereas set-valued occurrences specify the compact domain or support of a measure. In Theorem 3.3, Corollary 3.8, Proposition C.1, Proposition 4.1, Theorem 4.2, Proposition 4.3, Corollary 4.4: In the boundedness assumption, $K > 0$ satisfies $\max_{i \in [n_V], j \in [n_U]} \{\|V_i\|_\infty, \|U_j\|_\infty\} \leq K$; the indexed U_j are the latent variables separated from the categorical Gumbel noises. In Lemma E.1: $K \subset \mathbb{R}^{d_1}$ is the compact set carrying μ and $\widehat{\mu}$ and the common domain of F and $\widehat{F}$. In Lemma C.3: K bounds the regression functions: $|f(t)| \leq K$ and $|\widehat{f}(t)| \leq K$. In Proposition B.2, Corollary 3.8, Proposition C.1, Proposition 4.1, Theorem 4.2, Corollary 4.4, the result split from Theorem 3.6: In the lower-bound support assumption, $K \subset \mathbb{R}^D$ is compact and supports the measure under consideration; $f \in \mathcal{H}([0, 1]^d, K) \cap \mathcal{F}_L([0, 1]^d, K)$. This set-valued occurrence is distinct from a numerical boundedness constant. In Theorem 4.2: In the explanatory notation $\mathcal{P}(K)$, K is the set on which probability measures are defined; in the displayed factors $\mathcal{P}([-K, K]^{d_j^U})$, the K inside the cube endpoints is the numerical boundedness constant. In Theorem 4.2, Corollary 4.4: In the truncated neural classes, numerical K is the clipping level in $\max\{-K, \min\{f, K\}\}$ and is the boundedness constant chosen for the underlying SCM. Every true continuous structural function and every candidate continuous structural network additionally satisfies $\sup_x \|f_i(x)\|_1 \leq K$ and $\sup_x \|\widehat{f}_i(x)\|_1 \leq K$, respectively, on its full structural input domain. The limiting structural class

retains $\mathcal{F}_L^K = \{f : \sup_x \|f(x)\|_1 \leq K, \text{Lip}(f) \leq L\}$. In Lemma E.6: $K > 0$ is the uniform output bound $\|f(x)\|_\infty \leq K$ for every $f \in \mathcal{F}$ and every $x \in \mathbb{R}^{d'}$; consequently the outputs lie in $[-K, K]^d$. In Proposition C.1: For each non-Gumbel latent law $\mathbb{P}$ with support components C_i , Assumption 3.5 applies to every normalized component law $\mu_i = \mathbb{P}|_{C_i}/\mathbb{P}(C_i)$ with the compact support set $K = C_i$. In Proposition 4.1: For each non-Gumbel latent law $\mathbb{P}$ with support components C_i , Assumption 3.5 applies to every normalized component law $\mu_i = \mathbb{P}|_{C_i}/\mathbb{P}(C_i)$ with the compact support set $K = C_i$. In Corollary 3.8: For each non-Gumbel latent law $\mathbb{P} = P(U_C)$ with support components C_i , the lower-bound support assumption applies to each normalized component law $\mu_i = \mathbb{P}|_{C_i}/\mathbb{P}(C_i)$ with compact support set $K = C_i$. Its chart $f : [0, 1]^{d_i^C} \rightarrow C_i$ satisfies $(f^{-1})_\# \mu_i(\partial[0, 1]^{d_i^C}) = 0$ when $d_i^C > 0$. $\mathcal{F}_1$ is function-class notation whose subscript can be a class identifier or the Lipschitz constant in the previously defined notation $\mathcal{F}_L(K_1, K_2)$. In Corollary 3.8, Theorem 4.2: As the second argument of $\text{NCM}_G(\mathcal{F}_0, \mathcal{F}_1)$, the class restricts the latent generators, with input and output dimensions omitted from the notation. In Theorem 4.2: The latent-generator class is $\mathcal{F}_{1,n} = \mathcal{NN}^{\infty,K}(\Theta(d_{\max}^U), L_{1,n}, \Theta((d_{\max}^U)^2), L_{2,n}, \tau_n)$, a class varying with n . d_1 denotes a locally specified input or source dimension, including the input dimension in $\mathcal{NN}_{d_1,d_2}$, or an explicitly specified averaged metric on evaluation vectors. In Lemma E.1: $K \subset \mathbb{R}^{d_1}$ is the common domain supporting μ and $\hat{\mu}$. In Theorem 3.6: In the proof's notation for the combined maps $\hat{g}_i$, $d_1 = 1$. In a result cited in the proof of Lemma E.6: d_1 is the averaged metric on length- m evaluation vectors, $d_1(u, w) = \frac{1}{m} \sum_{i=1}^m |u_i - w_i|$, with m a positive integer. In a result cited in the proof of Lemma E.6: d_1 is the averaged metric on length- N evaluation vectors, $d_1(u, w) = \frac{1}{N} \sum_{i=1}^N |u_i - w_i|$; in the covering number appearing in the probability bound, $N = 2m$. d_2 denotes the locally specified output or target dimension of a map, including the output dimension in $\mathcal{NN}_{d_1,d_2}$. In Lemma E.1: $F, \hat{F} : K \rightarrow \mathbb{R}^{d_2}$, so d_2 is their common target dimension. In Theorem 3.6: In the proof's notation for the combined maps $\hat{g}_i$, $d_2 = d$. p_i denotes a nonnegative scalar weight indexed by the coordinates, mixture terms, or partition cells of the local construction. In Theorem 3.6: $p_i = \mathbb{P}(C_i)$ for $i \in [N_C]$, and $\sum_{i=1}^{N_C} p_i = 1$; these are the probabilities used to combine the support components. In Lemma E.2: $p_1, \dots, p_n \in [0, 1]$ are arbitrary mixture weights satisfying $\sum_{k=1}^n p_k = 1$, with no requirement that they be masses of connected support components. In Lemma E.4: $p_1, \dots, p_n > 0$ are arbitrary positive softmax weights; normalization is not required in the statement. The proof normalizes them without loss of generality, and the limiting categorical probabilities are $p_i / \sum_{k=1}^n p_k$. In Lemma E.5: In the probability-simplex notation, p_i is a coordinate satisfying $p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$. In Theorem 3.3: In the categorical case of the proof, $(p_1, \dots, p_{n_i}) = f(\text{pa}(v_k), \mathbf{u}_{V_k})$

is the conditional categorical probability vector evaluated at the displayed parent and latent values; its coordinates are nonnegative and sum to one. In Proposition B.1: In the probability-simplex notation, p_i is a coordinate satisfying $p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$. In the result split from Theorem 3.6: $p_i = \mathbb{P}(C_i)$ for $i \in [N_C]$, and $\sum_{i=1}^{N_C} p_i = 1$; these are the probabilities used to combine the support components. In Proposition B.2: In the printed partition construction at level $n + 1$, $p_i = \mathbb{P}(C_{O_{n+1}(i)})$ for $1 \leq i \leq 2^{(n+1)d}$. Here $\mathbb{P}$ is the reduced cube measure and $C_{O_{n+1}(i)}$ is an ordered partition cube. These masses give the lengths of the corresponding parameter intervals and the denominators in the polygonal curve; domination by Lebesgue measure makes them positive. In Theorem 4.2: In the latent approximation construction, p_i denotes the probability mass of component C_i under the particular latent target law being approximated. These component probabilities sum to one and weight the component approximation errors and the hard selected mixture. In the categorical measurability argument, p_i instead denotes coordinate i of the categorical propensity vector evaluated at the structural input; these coordinates are nonnegative and sum to one, and zero coordinates are allowed. Δ_n denotes a subset of $\mathbb{R}^n$ whose coordinates sum to one. In Lemma E.5: $\Delta_n = \{(p_1, \dots, p_n) \in \mathbb{R}^n : \sum_{i=1}^n p_i = 1, p_i \geq 0\}$ is the probability simplex. In Lemma E.4: In the proof, $\Delta_n = \{(x_1, \dots, x_n) \in \mathbb{R}^n : \sum_{k=1}^n x_k = 1\}$ is the domain of the transportation map; the vectors X^τ and X^0 themselves have nonnegative coordinates. In Theorem 3.3, Proposition B.1: $\Delta_n = \{(p_1, \dots, p_n) \in \mathbb{R}^n : \sum_{i=1}^n p_i = 1, p_i \geq 0\}$ is the probability simplex. B is a locally specified measurable test set or positive scalar bound. In the result split from Theorem 3.6: In the componentwise lower-bound premise, B ranges over measurable subsets of $[0, 1]^{d_i^C}$; the component measure and its parameterizing map satisfy the domination inequality of Assumption 3.5 on these sets. Separately, in the proof's clipping and mixing estimates, $B > 0$ is a fixed scalar chosen so that every support component lies in $[-B, B]^d$, and the second-block outputs are clipped into this box. In Proposition B.2, Corollary 3.8, Proposition C.1, Proposition 4.1, Theorem 4.2, Corollary 4.4: In the lower-bound premise, B ranges over measurable subsets of $[0, 1]^d$ and $\mathbb{P}(f(B)) \geq C_f \lambda(B)$; for a component with dimension d_i^C , the corresponding test sets lie in $[0, 1]^{d_i^C}$. In Proposition 4.3: B is the scalar bound $B = K(1 + 2qL_y)$, where $q = \sum_{i: V_i \text{ is continuous}} d_i + |\{i : V_i \text{ is categorical}\}|$ and $L_y = (LD + 1)^{n_V}$; it enters $C_W = 4L_y B + 2B \max\{L_y, 1\}$. In Theorem 3.3, Lemma E.1, Theorem 3.6, Lemma E.6, a result cited in the proof of Lemma E.6: In a push-forward expression $f_{\#}\mu(B)$, B is an arbitrary measurable subset of the codomain of f . In Theorem 3.3: In the proof's error recurrence, $B = \max(\{1\} \cup \{d_i : V_i \text{ is continuous}\} \cup \{n_i(n_i - 1)/2 : V_i \text{ is categorical}\})$. In Theorem 4.2: In the proof's model-comparison recurrence, B is a fixed positive bound on the coefficients of

the structural approximation errors, chosen together with $A \geq 1$ so that $e_{k+1} \leq Ae_k + Bd(\theta, \theta')$. Separately, in the derivation of the pseudo-dimension bound, $B = CH_n^2$, where $H_n = 1 + L_{1,n} + L_{2,n}$ and C is the constant in that derivation; this latter scalar may vary with n . In Corollary 3.8: In the structural extension and approximation construction, $B = B_0H^{n_V}$, where $m = d_{\max}^{\text{in}}$, $H = 1 + 2Lm$, and $B_0 = \max(\{1, R, K\} \cup \{n_i : V_i \text{ is categorical}\} \cup \{\|f_i(a_i)\|_\infty : V_i \text{ is continuous}\})$; a_i has zero in each real input coordinate and one in each categorical input coordinate. This fixed positive scalar is independent of the intervention and network budgets and defines the structural input cubes $Q_i = [-B, B]^{m_i}$. In Corollary 4.4: In the proof's model-comparison recurrence, B is a fixed positive bound on the coefficients of the structural approximation errors, chosen together with $A \geq 1$ so that $e_{k+1} \leq Ae_k + Bd(\theta, \theta')$; it is independent of the sample size and the compared model parameters. $\|\cdot\|_\infty$ denotes the infinity norm on vectors and its uniform extension to functions: $\|f\|_\infty = \sup_{x \in \text{dom}(f)} \|f(x)\|_\infty$. S denotes the auxiliary object defined locally in its surrounding argument. In Theorem 3.3: In the proof, $S = \sum_{i=1}^{n_V} \|f_i - \hat{f}_i\|_\infty$ is the summed structural-function approximation error. In Corollary 3.8: In the categorical normalization argument, at the structural input $x = (\text{pa}(v_k), \mathbf{u}_{V_k})$, $S = \|\hat{f}_i(x) - f_i(x)\|_\infty$, where $f_i(x)$ is the normalized true propensity vector and $\hat{f}_i(x)$ is the unnormalized nonnegative approximant. The comparison is with $\hat{f}_i(x)/\|\hat{f}_i(x)\|_1$ wherever its denominator is nonzero. In Lemma E.6, a result cited in the proof of Lemma E.6: S in $\mathcal{N}(\epsilon, S, \|\cdot\|_1)$ is the set to be covered; the unnormalized empirical construction uses $\{(f(x_1), \dots, f(x_N)) : f \in \mathcal{F}\}$ for specified input points. For $\mathcal{N}_1^{\text{avg}}(\epsilon, \mathcal{G}, m)$, the evaluation set is instead $\{(g(x_1), \dots, g(x_m)) : g \in \mathcal{G}\}$, equipped with $d_1(u, w) = m^{-1} \sum_{i=1}^m |u_i - w_i|$. In a result cited in the proof of Lemma E.6: The set covered in $\mathcal{N}_1^{\text{avg}}(r, \mathcal{G}, N)$ is $S = \{(g(x_1), \dots, g(x_N)) : g \in \mathcal{G}\}$ for specified evaluation points, equipped with $d_1(u, w) = N^{-1} \sum_{i=1}^N |u_i - w_i|$; the bound uses $N = 2m$.

3 Approximation Error of Neural Causal Models

In this section, we study the expressive power of NCMs, which serves as a key ingredient in proving the consistency result. In particular, given an SCM $\mathcal{M}^*$, we want to construct an NCM $\widehat{\mathcal{M}}$ such that the two causal models produce similar interventional results. Unlike in the discrete case [4, 50], latent distributions can be extremely complicated in general cases. The main challenge is how to design the structure of NCMs to ensure strong approximation power.

Remark (notation). $\widehat{\mathcal{M}}$ is a model on the same causal graph $\mathcal{G}$ as the model with which it is

compared; the hat alone does not denote a fitted estimator or an optimization minimizer. In Theorem 3.3: $\widehat{\mathcal{M}}$ is any comparison model in $\mathcal{M}(\mathcal{G}, \widehat{\mathcal{F}}, \widehat{\mathcal{U}})$; it is not required by this notation to be a neural model. In Corollary 3.8: $\widehat{\mathcal{M}}$ is the constructed neural causal model with the displayed structural and latent-generator architectures and structure equations 3.2, chosen after fixing $R > 0$; it may depend on R and the network budgets, but the same model applies to every intervention with $\|t_i\|_\infty \leq R$ for continuous treatment variables and $t_i \in [n_i]$ for categorical treatment variables. In Proposition 4.3: $\widehat{\mathcal{M}} \in \mathcal{M}(\mathcal{G}, \mathcal{F}_L, \mathcal{U})$ is the second comparison SCM, with observational law μ . In Proposition B.1: $\widehat{\mathcal{M}}$ is the canonical representation obtained by merging the original non-Gumbel latent variables and retaining the categorical Gumbel noises; its continuous structural functions are the original f_i under coordinate regrouping, and its categorical propensity functions are $f_i/\|f_i\|_1$ under the same regrouping, ignoring additional input coordinates, so its intervention laws equal those of $\mathcal{M}$. In Theorem 4.2, Corollary 4.4: In equation 4.1, $\widehat{\mathcal{M}}$ ranges over network representations in $\text{NCM}_{\mathcal{G}}(\mathcal{F}_{0,n}, \mathcal{F}_{1,n})$, with each categorical raw structural network q satisfying $q(x) \in [0, \infty)^{n_i}$, $\|q(x)\|_1 > 0$, and $\text{Lip}(x \mapsto q(x)/\|q(x)\|_1) \leq \widehat{L}_f$ on its full structural input domain. Different representations with the same latent population laws may have different common-draw empirical distributions. In Theorem 4.2: Every continuous structural network $\widehat{f}_i$ additionally satisfies $\sup_x \|\widehat{f}_i(x)\|_1 \leq K$ on its full structural input domain. In equation 4.2, $\widehat{\mathcal{M}}$ instead ranges over canonical SCMs in $\mathcal{M}(\mathcal{G}, \mathcal{F}_L^K, \mathcal{U})$ and all latent distributions $P(\mathcal{U})$, with $\mathcal{F}_L^K = \{f : \sup_x \|f(x)\|_1 \leq K, \text{Lip}(f) \leq L\}$ and normalized categorical propensity functions. In Corollary 4.4: Every continuous structural network $\widehat{f}_i$ additionally satisfies $\sup_x \|\widehat{f}_i(x)\|_1 \leq K$ on its full structural input domain. The population comparison models use the bounded Lipschitz structural class $\mathcal{F}_L^K = \{f : \sup_x \|f(x)\|_1 \leq K, \text{Lip}(f) \leq L\}$ and normalized categorical propensity functions.

In the following, we first derive an upper bound on the Wasserstein distance between two causal models sharing the same causal graph. Using this result, we decompose the approximation error into two parts: the error caused by approximating structural functions via neural networks and the error of approximating the latent distributions. Then, we design different architectures for these two parts.

Decomposing the Approximation Error First, we present a canonical representation of an SCM, which essentially states that we only need to consider the case where each latent variable U_i corresponds to a C^2 component of $\mathcal{G}$.

Definition 3.1 (Canonical representation). A SCM $\mathcal{M}$ with causal graph $\mathcal{G}$ has canonical form if

1. The set of latent variables consists of two sets,

$$U = \{U_C : C \text{ is a } C^2\text{-component of } \mathcal{G}\} \cup \{G_{V_i} = (g_1^{V_i}, \dots, g_{n_i}^{V_i}) : V_i \text{ is categorical}\},$$

where U_C and $g_j^{V_i}$ are independent and $g_j^{V_i}$ are standard Gumbel variables.

2. The structure equations have the form

$$V_i = \begin{cases} f_i(\text{Pa}(V_i), \mathbf{U}_{V_i}), & V_i \text{ is continuous,} \\ \arg \max_{k \in [n_i]} \{g_k^{V_i} + \log(f_i(\text{Pa}(V_i), \mathbf{U}_{V_i}))_k\}, & V_i \text{ is categorical, } \|\mathbf{f}_i\|_1 = 1, \end{cases} \quad (3.1)$$

where $\mathbf{U}_{V_i} = \{U_C : V_i \in C, C \text{ is a } C^2\text{-component of } \mathcal{G}\}$ and $(x)_k$ is the k -th coordinate of the vector x . We further assume that f_i are normalized for categorical variables.

Given a function class $\mathcal{F}$, the SCM class $\mathcal{M}(\mathcal{G}, \mathcal{F}, \mathbf{U})$ consists of all canonical SCM models with causal graph $\mathcal{G}$ such that $f_i \in \mathcal{F}, i \in [n_V]$.

Remark. A model belongs to $\mathcal{M}(\mathcal{G}, \mathcal{F}, \mathbf{U})$ when it is a canonical SCM whose structural equations use the latent variables only as the C^2 -components of $\mathcal{G}$ prescribe, with $f_i \in \mathcal{F}$; a bi-directed arrow of $\mathcal{G}$ permits shared latents but does not require them to be correlated, so the latent laws $P(U_C)$ may be arbitrary, degenerate laws included. In the categorical branch of (3.1), zero coordinates of the normalized nonnegative vector $f_i(x)$ are allowed and are evaluated using $\log 0 = -\infty$.

Remark (notation). U_C denotes the canonical latent vector indexed by a C^2 -component C of $\mathcal{G}$, and $\mathbf{U}_{V_i} = \{U_C : V_i \in C, C \text{ is a } C^2\text{-component of } \mathcal{G}\}$. In Corollary 3.8: For every C^2 -component C , the law $P(U_C)$ satisfies Assumption 3.4; for each connected support component C_i of positive mass, the normalized law $\mu_i = P(U_C)|_{C_i}/P(U_C \in C_i)$ satisfies Assumption 3.5 with $K = C_i$, and its support chart f satisfies $(f^{-1})_{\#}\mu_i(\partial[0, 1]^{d_i^C}) = 0$ when $d_i^C > 0$. $\mathcal{M}(\mathcal{G}, \mathcal{F}, \mathbf{U})$ consists of canonical SCMs on $\mathcal{G}$ with $f_i \in \mathcal{F}$ for $i \in [n_V]$ and the indicated canonical latent coordinates; the argument $\mathbf{U}$ does not fix a common latent probability law across the class.

Proposition B.1 in the appendix shows that any SCM satisfying ?? 2.3, 2.4 can be represented in this way and we provide an example in Section B.1 to illustrate how to obtain the canonical representation for a given SCM. Therefore, we restrict our attention to the class $\mathcal{M}(\mathcal{G}, \mathcal{F}, \mathbf{U})$. For two SCM classes $\mathcal{M}(\mathcal{G}, \mathcal{F}, \mathbf{U}), \mathcal{M}(\mathcal{G}, \widehat{\mathcal{F}}, \widehat{\mathbf{U}})$, we want to study how well we can represent the models in the first class by the second class. The Wasserstein distance between the intervention

distributions is used to measure the quality of the approximation. To approximate the functions in the structural equations, we need to make the following regularity assumptions on the functions.

Assumption 3.2. If V_i is continuous, f_i in (3.1) are L_f -Lipschitz continuous. If V_i is categorical, the propensity functions $f_i(\text{Pa}(V_i), \mathbf{U}_{V_i}) \triangleq \mathbb{P}(V_i = j | \text{Pa}(V_i), \mathbf{U}_{V_i}), j \in [n_i]$ are L_f -Lipschitz continuous. There exists a constant $K > 0$ such that $\max_{i \in [n_V], j \in [n_U]} \{\|V_i\|_\infty, \|U_j\|_\infty\} \leq K$.

Remark. For categorical V_i , $f_i(x) = (\mathbb{P}(V_i = j | x))_{j \in [n_i]}$ is the structural propensity vector, and $\|f_i(x) - f_i(x')\| \leq L_f \|x - x'\|$ on its structural input domain, using the paper's 1-norm. There is a deterministic $K > 0$ such that $\max_{i \in [n_V], j \in [n_U]} \{\|V_i\|_\infty, \|U_j\|_\infty\} \leq K$ almost surely under the SCM law.

The following theorem summarizes our approximation error decomposition.

Theorem 3.3. Given any SCM model $\mathcal{M} \in \mathcal{M}(\mathcal{G}, \mathcal{F}_L, \mathbf{U})$, let the treatment variable set be $\mathbf{T} = \{T_k\}_{k=1}^{n_t}$ and suppose that ?? 2.3, 2.4 and 3.2 hold for $\mathcal{M}$ with Lipschitz constant L , constant K . For any intervention $\mathbf{T} = \mathbf{t}$ and $\widehat{\mathcal{M}} \in \mathcal{M}(\mathcal{G}, \widehat{\mathcal{F}}, \widehat{\mathbf{U}})$, we have

$$W\left(P^{\mathcal{M}}(\mathbf{V}(\mathbf{t})), P^{\widehat{\mathcal{M}}}(\widehat{\mathbf{V}}(\mathbf{t}))\right) \leq C_{\mathcal{G}}(L, K; \{d_i, n_i\}) \left(\sum_{i=1}^{n_V} \|f_i - \widehat{f}_i\|_\infty + W\left(P^{\mathcal{M}}(\mathbf{U}), P^{\widehat{\mathcal{M}}}(\widehat{\mathbf{U}})\right) \right),$$

where $C_{\mathcal{G}}(L, K; \{d_i, n_i\})$ is a constant that only depends on L, K , the causal graph $\mathcal{G}$, the continuous output dimensions d_i , and the categorical cardinalities n_i , and $f_i, \widehat{f}_i$ are structural functions of $\mathcal{M}$ and $\widehat{\mathcal{M}}$ respectively.

Additional condition. In the categorical structural equations, $\arg \max$ selects the smallest maximizing label when there is a tie. This is a measurable completion on ties; under the stipulated independent standard Gumbel noises and nonnegative propensity weights of positive total, it leaves every fixed-intervention law unchanged.

Remark. For a categorical propensity coordinate equal to zero, $\log 0 = -\infty$ in the canonical $\arg \max$ construction. The structural uniform infinity norm takes the supremum of the coordinatewise infinity norm over the common structural input domain. Under $\mathbf{T} = \mathbf{t}$, every unreplaced equation is evaluated in topological order at the intervened values of its parents, with the original latent law unchanged; for canonical categorical nodes the unreplaced equation is the categorical branch of (3.1).

Remark (notation). $\widehat{\mathbf{U}}$ denotes the full canonical latent collection of $\widehat{\mathcal{M}}$, including categorical Gumbel noises, rather than only the non-Gumbel inputs to its structural functions. $\widehat{\mathbf{V}} = \{\widehat{V}_i\}_{i=1}^{n_V}$ is

the collection of observed variables of $\widehat{\mathcal{M}}$ corresponding to $\mathbf{V}$, and $\widehat{\mathbf{V}}(t)$ denotes this collection under intervention $\mathbf{T} = t$. For $i \in [n_V]$, $\widehat{f}_i$ is the comparison model's structural map for a continuous node and its categorical propensity or weight map for a categorical node, evaluated on parent values and the corresponding latent coordinates. In Theorem 3.3: The functions are those of $\widehat{\mathcal{M}} \in \mathcal{M}(\mathcal{G}, \widehat{\mathcal{F}}, \widehat{\mathcal{U}})$; categorical structural propensities are read under the canonical representation, including its normalization convention and $\log 0 = -\infty$ for zero coordinates. The quantity $\|f_i - \widehat{f}_i\|_\infty$ takes the supremum of the coordinatewise infinity norm over the common structural input domain. In Corollary 3.8, a result cited in the proof of Corollary 3.8: In the construction, let q_i be a minimizer of $\|q_i - f_i\|_\infty$ over $\mathcal{NN}_{d_i^{\text{in}}, d_i^{\text{out}}}(d_i^{\text{out}}(2d_i^{\text{in}} + 10), L_0)$. The chosen network is $\widehat{f}_i = q_i$ for a continuous node and $\widehat{f}_i = \max\{0, q_i\}$ coordinatewise for a categorical node; the latter's normalized propensity is $\widehat{f}_i / \|\widehat{f}_i\|_1$ wherever the denominator is nonzero. In Theorem 4.2: Continuous structural networks range over $\mathcal{F}_{0,n}$ subject additionally to $\sup_x \|\widehat{f}_i(x)\|_1 \leq K$ on the full structural input domain. For a categorical node, the raw network $q \in \mathcal{F}_{0,n}$ satisfies $q(x) \in [0, \infty)^{n_i}$ and $\|q(x)\|_1 > 0$ for every structural input x , and its normalized propensity satisfies $\text{Lip}(x \mapsto q(x)/\|q(x)\|_1) \leq \widehat{L}_f$; the proof's canonical structural-function coordinate is $q/\|q\|_1$. In Corollary 4.4: Continuous structural networks range over $\mathcal{F}_{0,n}$ subject additionally to $\sup_x \|\widehat{f}_i(x)\|_1 \leq K$ on the full structural input domain. For a categorical node, the raw network $q \in \mathcal{F}_{0,n}$ satisfies $q(x) \in [0, \infty)^{n_i}$ and $\|q(x)\|_1 > 0$ for every structural input x , and its normalized propensity satisfies $\text{Lip}(x \mapsto q(x)/\|q(x)\|_1) \leq \widehat{L}_f$. $\widehat{f}$ is a function on the domain of its paired function f , with the particular function and codomain specified by the local comparison. In Lemma C.3: $\widehat{f}(t) = \mathbb{E}_\mu[\widehat{Y} \mid \widehat{T} = t]$ is a population regression version, L -Lipschitz continuous and bounded by K , defined on the common treatment domain on which it is compared with f . $C_{\mathcal{G}}(L, K; \{d_i, n_i\})$ denotes a numerical comparison coefficient whose permitted dependence is only on $L, K, \mathcal{G}$, the continuous output dimensions d_i , and the categorical cardinalities n_i . $\widehat{\mathcal{F}}$ is the specified class containing the comparison model's structural functions on their respective node domains; categorical members are normalized propensity functions, and no neural architecture or Lipschitz bound is imposed merely by the hat.

Theorem 3.3 separates the approximation error into two parts, which motivates us to construct the NCM in the following way. First, we approximate the functions f_i in (3.1) by NNs $\widehat{f}_i$. Then, we approximate the distribution of latent variables by pushing forward uniform and Gumbel variables using neural networks, i.e., $\widehat{U}_{C_j} = \widehat{g}_j(Z_{C_j})$, where $\{C_j\}$ are C^2 components and Z_{C_j} are multi-variate uniform and Gumbel random variables. The structural equations of the resulting

approximated model $\widehat{\mathcal{M}}$ are

$$\widehat{V}_i = \begin{cases} \widehat{f}_i \left(\text{Pa}(\widehat{V}_i), (\widehat{g}_j(Z_{C_j}))_{U_{C_j} \in U_{V_i}} \right), & V_i \text{ is continuous,} \\ \arg \max_{k \in [n_i]} \left\{ g_k + \log \left(\widehat{f}_i \left(\text{Pa}(\widehat{V}_i), (\widehat{g}_j(Z_{C_j}))_{U_{C_j} \in U_{V_i}} \right) \right)_k \right\}, & V_i \text{ is categorical.} \end{cases} \quad (3.2)$$

For the first part, we need to approximate Lipschitz continuous functions. For simplicity, we assume that the domain of the functions are uniform cubes. Similar arguments hold for any bounded cubes. We denote $\mathcal{NN}_{k_1, k_2}(W, L)$ to be the set of ReLU NNs with input dimension k_1 , output dimension k_2 , width W and depth L . It has been shown that $\epsilon(\mathcal{NN}_{k_1, 1}(2d_1 + 10, L_0), \mathcal{F}_L([0, 1]^{d_1}, \mathbb{R})) \leq O(L_0^{-2/k_1})$ [52], where $\epsilon(\cdot, \cdot)$ denotes the approximation error defined in Section 2. For a vector valued function, we can use a wider NN to approximate each coordinate and get a similar rate.

Remark (notation). $\widehat{V}_i$ is the observed variable corresponding to V_i in $\widehat{\mathcal{M}}$, with $i \in [n_V]$ and the same continuous or categorical state space; $\widehat{V}_i(t)$ denotes its value after replacing treatment equations by $\mathbf{T} = \mathbf{t}$. In Corollary 3.8, Theorem 4.2: The structural equations are (3.2), with latent inputs $\widehat{U}_{C_j} = \widehat{g}_j(Z_{C_j})$ and structural functions $\widehat{f}_i$. In Theorem 3.3, Theorem 4.2: For a zero categorical propensity coordinate, the categorical argmax uses the extended logarithm $\log 0 = -\infty$. In Theorem 4.2: For a categorical node, the raw structural network q has nonnegative coordinates and $\|q(x)\|_1 > 0$ at every structural input, and its normalized propensity vector is $q(x)/\|q(x)\|_1$; ties in the categorical argmax are resolved by the smallest label. The two-argument notation $\mathcal{NN}_{k_1, k_2}(W, L)$ specifies a ReLU network class, with omitted dimensions supplied by context; the five-argument notation specifies the separable three-part generator architecture including its Gumbel-Softmax layer. In Corollary 3.8: The four-argument generator class $\mathcal{NN}(\Theta(d_{\max}^U), L_1, \Theta((d_{\max}^U)^2), L_2)$ suppresses the temperature τ of the same third Gumbel-Softmax layer, which remains present in the approximation bound. The multiplicative width factors in $\Theta(d_{\max}^{\text{in}} d_{\max}^{\text{out}})$, $\Theta(d_{\max}^U)$, and $\Theta((d_{\max}^U)^2)$ can be chosen as absolute numerical constants, independent of the causal model, R , all dimensions, and all category and component counts. In Theorem 3.6, the result split from Theorem 3.6: In the generator architecture, a component with $d_i^C = 0$ uses the unique map to $[0, 1]^0$ as its first deterministic block and a constant affine map to its singleton support component as its second deterministic block. In the result split from Theorem 3.6: The multiplicative width factors in $\Theta(D_C)$ and $\Theta(dD_C)$ are absolute numerical constants, independent of the measure, chart constants, dimensions, and component count. The notation $\widehat{g}_{\#}\lambda$ denotes the output law of the generator after integrating both its uniform input and its Gumbel-softmax randomness, rather than a push-forward conditional on a realized mixing vector. In Theorem 3.6: The generator has uniform-input dimension N_C , output dimension

d , and target law $\mathbb{P}$; its mixing weights are the target component masses. A component with $d_i^C = 0$ is a singleton, and its two deterministic blocks are the unique map to $[0, 1]^0$ and a constant affine map to that singleton. In Corollary 3.8, Theorem 4.2: In the NCM class notation, $\hat{g} \in \mathcal{F}_1$ is shorthand for the separate latent generators $\hat{g}_j$, whose compatible dimensions depend on the corresponding latent variable. In Theorem 4.2: At sample size n , generators range over $\mathcal{F}_{1,n}$ and have coordinatewise output truncation to $[-K, K]$, with finite component counts and component cube dimensions fixed as n varies. The optimization ranges over network representations, including distinct representations inducing the same latent population laws; each representation is evaluated on the same m_n exogenous input draws. In the result split from Theorem 3.6: The generator has uniform-input dimension N_C , output dimension d , and target law $\mathbb{P}$; its mixing weights are the target component masses. A component with $d_i^C = 0$ is a singleton, and its two deterministic blocks are the unique map to $[0, 1]^0$ and a constant affine map to that singleton. Z_{C_j} denotes the uniform and Gumbel input vector for the generator of the latent variable associated with graph C^2 -component C_j , with $N_{C,j}$ uniform coordinates and $N_{C,j}$ Gumbel coordinates, where $N_{C,j}$ counts the connected components of $\text{supp}(P(U_{C_j}))$. In Theorem 4.2: At each sample size n , the simulation inputs $Z_{C_j}^{(r)}$, $r \in [m_n]$, are shared across all candidate network representations: each candidate latent generator is evaluated as $\hat{g}_j(Z_{C_j}^{(r)})$. These inputs are drawn as part of the common independent and identically distributed exogenous product-law draws, independently of the observational sample. For a C^2 -component C_j of $\mathcal{G}$, $\hat{U}_{C_j} = \hat{g}_j(Z_{C_j})$ denotes the generated latent vector corresponding to U_{C_j} , with law $(\hat{g}_j)_\# P(Z_{C_j})$.

For the second part, we approximate each U_i individually by pushing forward i.i.d. multivariate uniform and Gumbel variables $\hat{U}_{C_i} = \hat{g}_i(Z_{C_i})$ since the latent variables are independent by ?? 2.3. To do so, we examine under what assumptions on the measure $\mathbb{P}$ over $\mathbb{R}^n$ we can find a NN $\hat{g}$ such that $W(\hat{g}_\# \lambda, \mathbb{P})$ is small, where $\lambda(\cdot)$ is some reference measure.

Remark (notation). $\mathbb{P}$ denotes the probability measure appropriate to the local distribution or events. In Proposition B.2: The statement uses $\mathbb{P}$ satisfying Assumption 3.5 on $K \subset \mathbb{R}^D$. The proof reduces to a measure on $[0, 1]^d$ that vanishes on the boundary and satisfies $\mathbb{P}(B) \geq C_1 \lambda(B)$ for all measurable $B \subset [0, 1]^d$; the reduced measure is $f_\#^{-1} \mathbb{P}$ for the homeomorphism in that assumption. Absolute continuity with respect to Lebesgue measure is assumed only in the illustrative proof sketch. In Theorem 3.6, a result cited in the proof of Theorem 3.6, the result split from Theorem 3.6: In the distribution-generator construction, $\mathbb{P}$ is the target probability measure on the ambient Euclidean space; expressions such as $\mathbb{P}(X_i = 1) = p_i$ in the selector construction denote probabilities under

the selector's law. In Theorem 3.3, Corollary 3.8, Proposition C.1, Proposition 4.1, Theorem 4.2, Proposition 4.3, Corollary 4.4: In the categorical propensity definition, $\mathbb{P}(V_i = j \mid \text{Pa}(V_i), \mathbf{U}_{V_i})$ denotes conditional probability under the relevant SCM law, rather than evaluation of the distribution-generator target measure. In a result cited in the proof of Theorem 3.6: $\mathbb{P}$ is the target probability measure on $\mathbb{R}^d$ satisfying Assumption 3.4; its normalized restrictions to the connected support components are the laws μ_i . In Proposition 4.1, Proposition C.1: In the componentwise lower-bound requirement, $\mathbb{P}$ denotes each non-Gumbel latent law; Assumption 3.5 is imposed on $\mathbb{P}|_{C_i}/\mathbb{P}(C_i)$ for every connected support component C_i , with $K = C_i$. C_i denotes a subset of the ambient Euclidean space in the local support or partition construction. In Theorem 3.6, Corollary 3.8, Proposition C.1, Proposition 4.1, Theorem 4.2, Corollary 4.4, a result cited in the proof of Theorem 3.6: C_i , $i \in [N_C]$, are the finitely many connected components of $\text{supp}(\mathbb{P})$, with $\text{supp}(\mathbb{P}) = \bigcup_{i=1}^{N_C} C_i$ and $\mathcal{H}([0, 1]^{d_i^C}, C_i) \cap \mathcal{F}_L([0, 1]^{d_i^C}, C_i) \neq \emptyset$ for some $d_i^C \geq 0$. In the result split from Theorem 3.6: C_i , $i \in [N_C]$, are the finitely many connected components of $\text{supp}(\mathbb{P})$, with $\text{supp}(\mathbb{P}) = \bigcup_{i=1}^{N_C} C_i$ and $\mathcal{H}([0, 1]^{d_i^C}, C_i) \cap \mathcal{F}_L([0, 1]^{d_i^C}, C_i) \neq \emptyset$ for some $d_i^C \geq 0$. In Theorem 3.6, the result split from Theorem 3.6: If $d_i^C = 0$, then C_i is a singleton, since it is homeomorphic to $[0, 1]^0$. In a result cited in the proof of Theorem 3.6: C_i , $i \in [N_C]$, are the finitely many connected components of $\text{supp}(\mathbb{P})$, with $\text{supp}(\mathbb{P}) = \bigcup_{i=1}^{N_C} C_i$ and $\mathcal{H}([0, 1]^{d_i^C}, C_i) \cap \mathcal{F}_L([0, 1]^{d_i^C}, C_i) \neq \emptyset$ for some $d_i^C \geq 0$; the quantized construction applies to components with $d_i^C > 0$. In Proposition 4.1, Proposition C.1: For each non-Gumbel latent law $\mathbb{P}$, every normalized component law $\mathbb{P}|_{C_i}/\mathbb{P}(C_i)$ satisfies Assumption 3.5 with compact support set $K = C_i$.

3.1 Approximating Mixed Distributions by Neural Networks

Remark (notation). N_C denotes the finite number of connected components of $\text{supp}(\mathbb{P})$, indexed by $i \in [N_C]$. In Theorem 4.2, Corollary 4.4: The finite component count of each latent generator architecture in $\mathcal{F}_{1,n}$ is fixed as n varies.

In this subsection, we will extend the results in [43] to construct a neural network as the push-forward map. It turns out that to get a good approximation of the targeted distribution, the shape of the support is essential. We introduce the following two assumptions for our latent distribution.

Assumption 3.4 (Mixed Distribution). The support of measure $\mathbb{P}$ has finite connected components $C_1, C_2 \dots C_{N_C}$, i.e., $\text{supp}(\mathbb{P}) = \bigcup_{i=1}^{N_C} C_i$, and each component C_i satisfies $\mathcal{H}([0, 1]^{d_i^C}, C_i) \cap \mathcal{F}_L([0, 1]^{d_i^C}, C_i) \neq \emptyset$ for some $d_i^C \geq 0$. Recall that $\mathcal{H}(K_1, K_2)$ is the set of homeomorphisms from

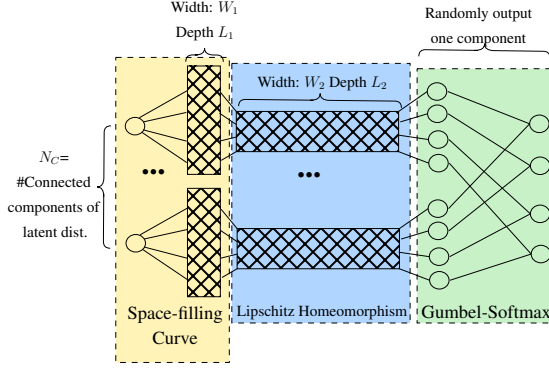


Figure 1: Architecture of neural network for 4-dimensional output. The first (yellow) part approximates the distribution on different connected components of the support using the results from [43]. The width and depth of each block in this part are W_1 and L_1 . The second (blue) part transforms the distributions on the unit cube to the distributions on the support. The width and depth of each block in the blue part are W_2 and L_2 . The third (green) part is the Gumbel-Softmax layer. It combines the distributions on different connected components of the support together and outputs the final distribution.

K_1 to K_2 defined at the end of Section 2.

Remark (notation). For $i \in [N_C]$, d_i^C is a nonnegative integer such that $\mathcal{H}([0, 1]^{d_i^C}, C_i) \cap \mathcal{F}_L([0, 1]^{d_i^C}, C_i) \neq \emptyset$; $d_i^C = 0$ corresponds to a singleton component. In Theorem 4.2, Corollary 4.4: The component cube dimensions of each latent generator architecture in $\mathcal{F}_{1,n}$ are fixed as n varies.

?? 3.4 encompasses almost all natural distributions. For example, distributions supported on $[0, 1]^d$ and closed balls, finitely supported distributions and mixtures of them all satisfy this assumption. ?? 3.4 allows us to transform the support of the targeted distribution into unit cubes and the nice geometric properties of unit cubes facilitate our construction. The next assumption is slightly than ?? 3.4, but we will show later that it can give us a better rate.



Figure 2: This figure demonstrates the first two parts of our architecture. Each interval in the yellow box corresponds to one coordinate of input in the left figure. We first push forward uniform distributions to different cubes. Then, using ?? 3.4, we adapt the shape of the support and push the measure from unit cubes to the original support of $P(U)$. In this way, we can approximate complicated measures by pushing forward uniform variables.

Assumption 3.5 (Lower Bound). Suppose that $\mathbb{P}$ is supported on a compact set $K \subset \mathbb{R}^D$, there exists a constant $C_f > 0$, $f \in \mathcal{H}([0, 1]^d, K) \cap \mathcal{F}_L([0, 1]^d, K)$, such that for any measurable set $B \subset [0, 1]^d$, $\mathbb{P}(f(B)) \geq C_f \lambda(B)$. Besides, if $d > 0$, $(f^{-1})_{\#} \mathbb{P}$ vanishes on the boundary $\partial[0, 1]^d$.

Remark (notation). C_f is a positive domination constant associated with the particular measure $\mathbb{P}$ and chosen homeomorphism f in the lower-bound assumption.

?? 3.5 implies that $d\lambda/d(f_{\#}^{-1}\mathbb{P})$ exists and is lowered bounded by a constant C_f . We show in the appendix (Proposition B.2) that under ?? 3.5, it is possible to simulate any distribution on the unit cubes with Hölder continuous curves and uniform distribution on $[0, 1]$. Proposition B.2 can be viewed as an extension of the Skorohod representation theorem.

Now, we briefly explain the construction of the push-forward maps. An example is provided in Section B.1 to illustrate the construction. The NN architecture consists of three parts (see Figure 1). The input dimension is the same as the number of connected components of the support N_C . For each component C_i , let $H_i \in \mathcal{H}([0, 1]^{d_i^C}, C_i) \cap \mathcal{F}_L([0, 1]^{d_i^C}, C_i)$, where d_i^C is the dimension of component C_i in ?? 3.4. By $\mathbb{P} = (H_i)_{\#}(H_i^{-1})_{\#} \mathbb{P}$ on C_i , we can approximate $(H_i^{-1})_{\#} \mathbb{P}$ first, which is supported on a unit cube. [43] constructs a wide NN $\hat{g}$ of width W and constant depth such that $W(\hat{g}_{\#} \lambda, (H_i)_{\#}^{-1} \mathbb{P}) \leq O(W^{-1/d_i^C})$ where λ is the uniform measure on $[0, 1]$. Then, we approximate the Lipschitz map H_i to C_i to pull the distribution back to C_i . These are the first two parts (yellow and blue blocks in Figure 1) of the architecture.

Suppose that the output of i -th coordinate in the first two parts is $\vec{v}_i$, the Gumbel-softmax layer in the third part (green box) combines different components of the support. In particular, we want to output $\vec{v}_i$ with probability $p_i = \mathbb{P}(C_i)$. Let $V = [\vec{v}_1, \dots, \vec{v}_i]$, this can be achieved by outputting VX , where $X = (X_1, \dots, X_{N_C})^T$ is a one-hot random vector with $\mathbb{P}(X_i = 1) = p_i$. To use backpropagation in training, we use the Gumbel-Softmax trick [31] to (approximately) simulate such a random vector, $\hat{X}_i^{\tau} = \frac{\exp((\log p_i + G_i)/\tau)}{\sum_{k=1}^{N_C} \exp((\log p_k + G_k)/\tau)}$, where $\tau > 0$ is the temperature parameter (a hyperparameter) and $G_i \sim \text{Gumbel}(0, 1)$ are i.i.d. standard Gumbel variables. As $\tau \rightarrow 0$, the distribution of $\hat{X}^{\tau}$ converges almost surely to the categorical distribution [36, Proposition 1]. In particular, when $\tau = 0$, we denote $\hat{X}_i^0 = \mathbb{I}_{i=\arg \max_j \{\log p_j + G_j\}}$. The output of the last layer is $V \hat{X}^{\tau}$. Note that the Gumbel-softmax function is differentiable with respect to parameter $\{\log(p_i)\}_{i=1, \dots, N_C}$. Therefore, we can train the network with common gradient-based algorithms.

Remark (notation). G_k is the k th standard Gumbel coordinate, with density $x \mapsto \exp(-x - \exp(-x))$ on $\mathbb{R}$, in the relevant Gumbel-softmax vector. In Lemma E.4: The coordinate index is $k \in [n]$. In

Theorem 3.6: $G_k, k \in [N_C]$, are the independent standard Gumbel coordinates of the mixing layer, one per connected support component, independent of the uniform inputs used by the deterministic blocks; their randomness is included in $\hat{g}_\# \lambda$. In the result split from Theorem 3.6: $G_k, k \in [N_C]$, are the independent standard Gumbel coordinates of the mixing layer, one per connected support component, independent of the uniform inputs used by the deterministic blocks; their randomness is included in $\hat{g}_\# \lambda$. $\tau > 0$ is the Gumbel-softmax temperature hyperparameter; a superscript 0 on a softmax output or layer denotes the limit as $\tau \rightarrow 0$, rather than evaluation of the positive-temperature formula at zero. In Lemma E.4: The temperature satisfies $0 < \tau < 1$. In Theorem 3.6: The error estimate is asymptotic as the network budgets tend to infinity and $\tau \downarrow 0$, with constants independent of the budgets and τ ; the proof uses the Gumbel-softmax estimate for $0 < \tau < 1$. In Theorem 4.2: The temperature is the positive sequence τ_n appearing in $\mathcal{F}_{1,n}$, with $\tau_n \rightarrow 0$. In the result split from Theorem 3.6: The error estimate is asymptotic as the network budgets tend to infinity and $\tau \downarrow 0$, with constants independent of the budgets and τ ; the proof uses the Gumbel-softmax estimate for $0 < \tau < 1$. In Corollary 4.4: The generator temperature is the positive sequence τ_n in $\mathcal{F}_{1,n}$ and in the approximation term $\tau_n(1 - \log \tau_n)$. In Corollary 3.8: The displayed error bound is asymptotic as $L_0, L_1, L_2 \rightarrow \infty$ and $\tau \downarrow 0$, with $R > 0$ and the causal model fixed. The implicit constant is independent of the intervention and the network budgets, but may depend on R and the fixed causal model.

In summary, our construction has the form $\hat{g} = \hat{g}_3^\tau \circ \hat{g}_2 \circ \hat{g}_1$, where $\hat{g}_i, i = 1, 2$ have the separable form $(\hat{g}_i^1(x_1), \dots, \hat{g}_i^{N_C}(x_{N_C}))$ and $\hat{g}_1^j \in \mathcal{NN}_{1,d_j^C}(W_1, L_1)$, $\hat{g}_2^j \in \mathcal{NN}_{d_j^C,d}(W_2, L_2)$, $j \in [N_C]$; $\{d_j^C\}$ are the dimension of cubes in ?? 3.4. $\hat{g}_3^\tau$ is the Gumbel-Softmax layer with temperature parameter $\tau > 0$. We slightly abuse the notations and denote this class of neural nets by $\mathcal{NN}(W_1, L_1, W_2, L_2, \tau)$. Putting things together, we obtain the following theorem.

Theorem 3.6. *Given any probability measure $\mathbb{P}$ on $\mathbb{R}^d$ that satisfies ?? 3.4, let λ be the Lebesgue measure on $[0, 1]^{N_C}$, where N_C is defined in ?? 3.4. Write $D_C = \max\{1, \max_i d_i^C\}$.*

1. *There exists a neural network $\hat{g} \in \mathcal{NN}(W_1, \Theta(D_C), \Theta(d \cdot D_C), L_2, \tau)$ with the above architecture (Figure 1) such that*

$$W(\hat{g}_\# \lambda, \mathbb{P}) \leq O \left(\mathbb{I}_{\{\max_i d_i^C > 0\}} \left(W_1^{-1/D_C} + L_2^{-2/D_C} \right) + (\tau - \tau \log \tau) \right).$$

Remark. A network in $\mathcal{NN}_{k_1,k_2}(W, L)$ has L hidden layers, each of width at most W , followed by the affine output layer (width at most W and exactly W give the same class of functions: a narrower

layer is padded with zero units). The covering number $\mathcal{N}(\epsilon, S, \|\cdot\|_1)$ is that of the cited source (Anthony and Bartlett): the least size of a subset $C \subseteq S$ such that every point of S lies at ℓ_1 distance strictly less than ϵ from some point of C . The law $\hat{g}_\# \lambda$ includes the independent standard Gumbel variables used in the mixing layer. The error estimates are asymptotic as the network budgets tend to infinity and $\tau \downarrow 0$, with constants independent of these budgets and of τ . A component with $d_i^C = 0$ is a singleton, and its two deterministic blocks are the unique map to $[0, 1]^0$ and a constant affine map to that singleton. In the proposition entitled “Deep distribution approximation”, the multiplicative width factors in $\Theta(D_C)$ and $\Theta(dD_C)$ are absolute numerical constants, independent of the measure, chart constants, dimensions, and component count. $f \in \mathcal{H}(K_1, K_2)$ means that $f : K_1 \rightarrow K_2$ is bijective and both f and $f^{-1} : K_2 \rightarrow K_1$ are continuous. A network in $\mathcal{NN}_{k_1, k_2}(W, L)$ has L hidden ReLU layers of width at most W , followed by an affine output layer of dimension k_2 . The layer uses $V = [\vec{v}_1, \dots, \vec{v}_{N_C}]$ and the same Gumbel draws for all $\tau > 0$, with $\lim_{\tau \downarrow 0} \hat{X}_i^\tau = \hat{X}_i^0 = \mathbb{I}_{\{i = \arg \max_{j \in [N_C]} (\log p_j + G_j)\}}$ almost surely for every $i \in [N_C]$. For each component of positive dimension $q = d_i^C$, the application of Theorem VIII.1 of [43] uses integer depth parameter $s = 1$, quantization parameter $\delta = 1/\lceil \sqrt{q}(q+1)n(n-1) \rceil$, and quantization mesh δ/n . The refinement integer is $n = \lfloor (W_1/A_q)^{1/q} \rfloor$ for $W_1 \geq A_q 2^q$, where $A_q \geq 1$ depends only on q and bounds the clipped network width by $A_q n^q$.

Remark (notation). This is the three-part generator class with component count N_C , component cube dimensions d_j^C , and output dimension d suppressed from its arguments; W_1, L_1 and W_2, L_2 specify the respective component-network widths and depths, and $\tau > 0$ specifies the mixing temperature. In Corollary 3.8: The four-argument generator-class occurrence suppresses the temperature argument; the latent generators retain the Gumbel-softmax temperature τ appearing in the bound. The multiplicative width factors in $\Theta(d_{\max}^U)$ and $\Theta((d_{\max}^U)^2)$ can be chosen as absolute numerical constants, independent of the causal model, R , all dimensions, and all category and component counts. In Theorem 4.2: The class used is its truncated version $\mathcal{NN}^{\infty, K}(\Theta(d_{\max}^U), L_{1,n}, \Theta((d_{\max}^U)^2), L_{2,n}, \tau_n)$; the superscript ∞ imposes no finite Lipschitz bound, while K specifies coordinatewise output truncation, and the finite component counts and component cube dimensions of the latent generator architectures are fixed as n varies. In Theorem 3.6, the result split from Theorem 3.6: The architecture includes components with $d_i^C = 0$: their first deterministic block is the unique map to $[0, 1]^0$, and their second deterministic block is a constant affine map to the corresponding singleton. In Corollary 4.4: The latent generator class is the truncated class $\mathcal{F}_{1,n} = \mathcal{NN}^{\infty, K}(\Theta(d_{\max}^U), L_{1,n}, \Theta((d_{\max}^U)^2), L_{2,n}, \tau_n)$; the superscript ∞ imposes no finite Lipschitz bound, K specifies coordinatewise output truncation,

and the finite component counts and component cube dimensions are fixed as n varies. In the result split from Theorem 3.6: The multiplicative width factors in $\Theta(D_C)$ and $\Theta(dD_C)$ are absolute numerical constants, independent of the measure, chart constants, dimensions, and component count. Each coordinate block $\widehat{g}_1^j$ uses one scalar input to generate coordinates in the cube $[0, 1]^{d_j^C}$ associated with component j , without mixing inputs from different components. In Theorem 3.6, the result split from Theorem 3.6: Writing $\mu_j = \mathbb{P}_{C_j}/\mathbb{P}(C_j)$, for $d_j^C \geq 1$ the first block approximates the cube law $((g_2^j)^{-1})_{\#}\mu_j$ by the law of $\widehat{g}_1^j(U_j)$; for $d_j^C = 0$, it is the unique map to $[0, 1]^0$. Each coordinate block $\widehat{g}_2^j$ maps the d_j^C -dimensional cube coordinates for component j into $\mathbb{R}^d$, without mixing inputs from different components. In Theorem 3.6, the result split from Theorem 3.6: For $d_j^C \geq 1$, the block $\widehat{g}_2^j$ approximates a map $g_2^j \in \mathcal{H}([0, 1]^{d_j^C}, C_j) \cap \mathcal{F}_L([0, 1]^{d_j^C}, C_j)$; for $d_j^C = 0$, it is a constant affine map to the singleton C_j . $\widehat{g}_3^\tau$ mixes the N_C component outputs using the Gumbel-softmax coefficient vector, with $\tau > 0$; the notation $\widehat{g}_3^0$ denotes its almost-sure hard-selection limit as $\tau \rightarrow 0$. In Theorem 3.6: The mixing weights are $p_k = \mathbb{P}(C_k)$ for the connected components C_k of the target support. In Corollary 3.8: The layer is used separately for each latent generator, with mixing weights given by the masses of the connected components of that latent variable's target law. In Theorem 4.2: At sample size n , the generator class uses temperature $\tau_n > 0$ and truncates the resulting generator output coordinatewise to $[-K, K]$. In the result split from Theorem 3.6: The mixing weights are $p_k = \mathbb{P}(C_k)$ for the connected components C_k of the target support.

Corollary 3.7 (Deep distribution approximation). *Given any probability measure $\mathbb{P}$ on $\mathbb{R}^d$ that satisfies ?? 3.4, let λ be the Lebesgue measure on $[0, 1]^{N_C}$, where N_C is defined in ?? 3.4, and write $D_C = \max\{1, \max_i d_i^C\}$. Suppose in addition that $\mathbb{P}$ constrained to each component satisfies ?? 3.5. There exists a neural network $\widehat{g} \in \mathcal{NN}(\Theta(D_C), L_1, \Theta(d \cdot D_C), L_2, \tau)$ with the architecture of Figure 1 such that*

$$W(\widehat{g}_{\#}\lambda, \mathbb{P}) \leq O\left(\mathbb{I}_{\{\max_i d_i^C > 0\}}\left(L_1^{-2/D_C} + L_2^{-2/D_C}\right) + (\tau - \tau \log \tau)\right).$$

Note that $\widehat{g}_3^\tau$ (the Gumbel-softmax layer) is actually a random function since the coefficient vector $\widehat{X}^\tau$ is a random variable. In this sense, $\widehat{g}_{\#}\lambda$ can be viewed as pushing forward uniform and Gumbel variables using a neural net.

The second conclusion of Theorem 3.6 shows that under one additional assumption on the distribution, deep ReLU networks have a stronger approximation power in approximating distributions, which means we can use fewer computational units to achieve the same worst-case theoretical

approximation error.

3.2 NCM Construction

Now, we can use the previous distribution approximation results to construct NCM approximations. For simplicity, we omit the input and output dimensions of the neural network. As we mention in previous sections, our construction (3.2) consists of two parts, $\hat{f}_i$ approximating the structural functions f_i in (3.1), and $\hat{g}_j(Z_j)$ approximating the latent variables U_j . Let $\text{NCM}_{\mathcal{G}}(\mathcal{F}_0, \mathcal{F}_1)$ be a collection of NCMs with structural equation (3.2) and $\hat{f}_i \in \mathcal{F}_0, \hat{g} \in \mathcal{F}_1$. By combining Theorem 3.3 and Theorem 3.6, we have the following corollary.

Remark (notation). $\text{NCM}_{\mathcal{G}}(\mathcal{F}_0, \mathcal{F}_1)$ is the collection of models on $\mathcal{G}$ with structural equations (3.2), in which every structural network belongs to $\mathcal{F}_0$ and every component-specific latent generator belongs to $\mathcal{F}_1$, with compatible dimensions suppressed. In Theorem 4.2, Corollary 4.4: At sample size n , the candidate class is $\text{NCM}_{\mathcal{G}}(\mathcal{F}_{0,n}, \mathcal{F}_{1,n})$, with the architecture and truncation specified in Theorem 4.2, restricted so that every categorical raw structural network q satisfies $q(x) \in [0, \infty)^{n_i}$, $\|q(x)\|_1 > 0$ for every structural input x , and $\text{Lip}(x \mapsto q(x)/\|q(x)\|_1) \leq \hat{L}_f$. In Theorem 4.2: The empirical optimization ranges over network representations in the restricted candidate class; representations with the same latent population laws remain distinct candidates and may yield different common-draw empirical distributions. Every admitted continuous structural network additionally satisfies $\sup_x \|\hat{f}_i(x)\|_1 \leq K$ on its full structural input domain. In Corollary 4.4: Candidates are evaluated as network representations on the same m_n exogenous simulation draws; representations with identical latent population laws can yield different empirical distributions on these common draws. Every admitted continuous structural network additionally satisfies $\sup_x \|\hat{f}_i(x)\|_1 \leq K$ on its full structural input domain. $\mathcal{F}_0$ is the structural-network class supplied to $\text{NCM}_{\mathcal{G}}(\mathcal{F}_0, \mathcal{F}_1)$, with the input and output dimensions appropriate to each node omitted from the notation. In Theorem 4.2: The class is $\mathcal{F}_{0,n} = \mathcal{NN}^{\hat{L}_f, K}(W_{0,n}, \Theta(\log d_{\max}^{\text{in}}))$, where $\hat{L}_f = \sqrt{d_{\max}^{\text{in}} d_{\max}^{\text{out}}} L_f$; it consists of coordinatewise truncations to $[-K, K]$ of networks whose Lipschitz constant is at most $\hat{L}_f$, and varies with n . When $d_{\max}^{\text{in}} = 1$, the depth expression uses $\Theta(1 + \log d_{\max}^{\text{in}})$. Continuous structural networks admitted to the optimization additionally satisfy $\sup_x \|\hat{f}_i(x)\|_1 \leq K$ on their full structural input domains. Categorical raw networks admitted to the optimization additionally satisfy $q(x) \in [0, \infty)^{n_i}$, $\|q(x)\|_1 > 0$ for every structural input x , and $\text{Lip}(x \mapsto q(x)/\|q(x)\|_1) \leq \hat{L}_f$.

Corollary 3.8. *Given a causal model $\mathcal{M}^* \in \mathcal{M}(\mathcal{G}, \mathcal{F}_L, \mathbf{U})$, suppose that Assumptions 2.3-3.2 hold and the distributions of U_C for all C^2 -component satisfy the assumptions of the proposition entitled “Deep distribution approximation”. Let $d_{\max}^{\text{in}}$ and $d_{\max}^{\text{out}}$ be the largest input and output dimension of f_i in (3.1) and $d_{\max}^U$ be the largest dimension of all latent variables. Suppose in addition that $d_{\max}^{\text{in}} > 0$ and $d_{\max}^U > 0$. Fix any $R > 0$. There exists a neural causal model*

$$\widehat{\mathcal{M}} \in \text{NCM}_{\mathcal{G}}(\mathcal{NN}(\Theta(d_{\max}^{\text{in}} d_{\max}^{\text{out}}), L_0), \mathcal{NN}(\Theta(d_{\max}^U), L_1, \Theta((d_{\max}^U)^2), L_2))$$

with structure equations (3.2). For any intervention $\mathbf{T} = \mathbf{t}$ such that $\|t_i\|_{\infty} \leq R$ for each continuous treatment variable T_i , $\widehat{\mathcal{M}}$ satisfies

$$W\left(P^{\mathcal{M}^*}(\mathbf{V}(\mathbf{t})), P^{\widehat{\mathcal{M}}}(\mathbf{V}(\mathbf{t}))\right) \leq O\left(L_0^{-2/d_{\max}^{\text{in}}} + L_1^{-2/d_{\max}^U} + L_2^{-2/d_{\max}^U} + (\tau - \tau \log \tau)\right).$$

The implicit constant is independent of the intervention and the network budgets, but may depend on R and the fixed causal model.

Additional condition. In the categorical structural equations, $\arg \max$ selects the smallest maximizing label when there is a tie. This is a measurable completion on ties; under the stipulated independent standard Gumbel noises and nonnegative propensity weights of positive total, it leaves every fixed-intervention law unchanged.

Remark. In Corollary 3.8, for $m_i = d_i^{\text{in}} > 0$, the symbol $\widehat{f}_i$ denotes $q_i \circ \Pi_i$ for a continuous variable and the vector $(\max\{\epsilon_i, (q_i \circ \Pi_i)_k\})_{k=1}^{n_i}$ for a categorical variable, where q_i , Π_i , and $\epsilon_i = A_i L_0^{-2/m_i}$ are chosen in its proof. In Corollary 3.8, for $m_i = 0$, the symbol $\widehat{f}_i$ denotes the exact constant function f_i . The error estimate is asymptotic as $L_0, L_1, L_2 \rightarrow \infty$ and $\tau \downarrow 0$, for fixed R and fixed causal model. The latent generator class in the statement uses the same temperature τ as the displayed error estimate. The multiplicative width factors in $\Theta(d_{\max}^{\text{in}} d_{\max}^{\text{out}})$, $\Theta(d_{\max}^U)$, and $\Theta((d_{\max}^U)^2)$ can be chosen as absolute numerical constants, independent of the causal model, R , all dimensions, and all category and component counts. For each latent law $P(U_C)$, the assumptions of the proposition entitled “Deep distribution approximation” require the support charts of ?? 3.4 and ?? 3.5 on each normalized component law $\mu_i = P(U_C)|_{C_i}/P(U_C \in C_i)$, with $(f^{-1})_{\#} \mu_i(\partial[0, 1]^{d_i^C}) = 0$ when $d_i^C > 0$. The phrase “uniform cubes” denotes the unit cube domains $[0, 1]^m$ used for deterministic uniform function approximation; it imposes no uniform probability law on structural inputs. For each categorical V_i , the noise in equation (3.2) satisfies $g_k = g_k^{V_i}$ for $k \in [n_i]$. Membership in $\text{NCM}_{\mathcal{G}}(\mathcal{F}_0, \mathcal{F}_1)$ requires $\widehat{f}_i \in \mathcal{F}_0$ for every observed node and

$\hat{g}_j \in \mathcal{F}_1$ for every latent component, with the compatible input and output dimensions understood separately for each network. $d_{\max}^U = \max_{j \in [n_U]} d_j^U$ ranges over non-Gumbel latent variables, while $d_{\max}^{\text{in}} = \max_i d_i^{\text{in}}$ and $d_{\max}^{\text{out}} = \max_i d_i^{\text{out}}$, with $d_i^{\text{out}} = n_i$ for categorical nodes.

Remark (notation). $d_{\max}^{\text{in}} = \max_{i \in [n_V]} d_i^{\text{in}}$, where d_i^{in} is the input dimension of $f_i(\text{Pa}(V_i), U_{V_i})$ in the canonical structural equations. In Theorem 4.2, Corollary 4.4: $d_{\max}^{\text{in}} \geq 1$; when $d_{\max}^{\text{in}} = 1$, the structural-network depth expression $\Theta(\log d_{\max}^{\text{in}})$ means $\Theta(1 + \log d_{\max}^{\text{in}})$. In Corollary 3.8: $d_{\max}^{\text{in}} > 0$. $d_{\max}^{\text{out}} = \max_{i \in [n_V]} d_i^{\text{out}}$, where $d_i^{\text{out}} = d_i$ for a continuous structural function and $d_i^{\text{out}} = n_i$ for a categorical propensity vector. $d_{\max}^U = \max_{j \in [n_U]} d_j^U$, where d_j^U is the dimension of a canonical latent variable U_j whose distribution is approximated by a generator, distinct from the separately supplied categorical Gumbel noises. In Theorem 4.2, Corollary 4.4: $d_{\max}^U \geq 1$; it remains the maximum individual non-Gumbel latent dimension, distinct from the total dimension $D_U = \sum_{j=1}^{n_U} d_j^U$ and $D = \max\{1, D_U\}$. In Corollary 3.8: $d_{\max}^U > 0$.

Similar approximation results also hold for wide NN approximation (the first construction in Theorem 3.6), as presented in Section 3.1. The proof can be easily adapted to the wide NNs architecture. Corollary 4.4 gives an explicit construction of NCM for a given SCM, which shows the expressivity of NCMs for SCM with mixed variables. This result can be viewed as a generalization of the expressivity result in [51], but our construction is completely different from [51].

4 Consistency of Neural Causal Partial Identification

In this section, we prove the consistency of the max/min optimization approach to partial identification. In the finite sample setting, we consider the following estimator F_n of the optimal values of (1.1).

$$F_n = \underset{\widehat{\mathcal{M}} \in \text{NCM}_{\mathcal{G}}(\mathcal{F}_{0,n}, \mathcal{F}_{1,n})}{\text{argmin}} \quad \mathbb{E}_{t \sim \mu_T} \mathbb{E}_{\widehat{\mathcal{M}}} [F(V_1(t), \dots, V_{n_V}(t))], \quad (4.1)$$

$$s.t. \quad S_{\lambda_n}(P_{m_n}^{\widehat{\mathcal{M}}}(\mathbf{V}), P_n^{\mathcal{M}^*}(\mathbf{V})) \leq \alpha_n,$$

where $P_n^{\mathcal{M}^*}, P_{m_n}^{\widehat{\mathcal{M}}}$ are the empirical distribution of $P^{\mathcal{M}^*}, P^{\widehat{\mathcal{M}}}$ with sample size n, m_n , μ_T is a given measure, $S_{\lambda}(\cdot, \cdot)$ is the Sinkhorn distance and $\mathcal{F}_{i,n}$ will be specified later. For example, the counterfactual outcome $\mathbb{E}[Y(1)]$ is a special case of the objective. Our results can be easily generalized to any linear combination of objective functions of this form. We use the Sinkhorn distance because it can be computed efficiently in practice [15]. We want to study if F_n is a good

lower bound as $n \rightarrow \infty$.

To match the observational distribution, we need to increase the width or depth of the NNs we use. As the sample size increases, the number of parameters also increases to infinity, which creates difficulty for our analysis. To obtain consistency, we need to regularize the functions while preserving their approximation power. Surprisingly, if we do not use any regularization, the following proposition implies that consistency may not hold even if the SCM is identifiable.

Proposition 4.1 (Informal, see Proposition C.1 for a formal version). *There exists a constant $c > 0$ and an identifiable SCM $\mathcal{M}^*$ satisfying Assumptions 2.3-3.5 such that for any $\epsilon > 0$, there exists an SCM $\mathcal{M}_\epsilon$ satisfying $W(P^{\mathcal{M}^*}(\mathbf{V}), P^{\mathcal{M}_\epsilon}(\mathbf{V})) \leq \epsilon$ and $|\text{ATE}_{\mathcal{M}^*} - \text{ATE}_{\mathcal{M}_\epsilon}| > c$. Here, for each non-Gumbel latent law $\mathbb{P}$ with connected support components C_i , Assumption 3.5 is required of each normalized component law $\mathbb{P}|_{C_i}/\mathbb{P}(C_i)$, with $K = C_i$.*

Additional condition. In the categorical structural equations, $\arg \max$ selects the smallest maximizing label when there is a tie. This is a measurable completion on ties; under the stipulated independent standard Gumbel noises and nonnegative propensity weights of positive total, it leaves every fixed-intervention law unchanged.

Remark. Identifiable means: every SCM $\mathcal{M}$ on the backdoor graph of Proposition C.1 whose latent variables are all independent (Assumption 2.3) and with $P^{\mathcal{M}}(\mathbf{V}) = P^{\mathcal{M}^*}(\mathbf{V})$ has $\text{ATE}_{\mathcal{M}} = \text{ATE}_{\mathcal{M}^*}$; the comparison models $\mathcal{M}_\epsilon$ have that same graph. Models whose latents are pairwise independent but jointly dependent are not on the graph in this sense.

Remark (notation). $\mathcal{M}_\epsilon$ denotes an SCM chosen for each tolerance $\epsilon > 0$ in the counterexample family, rather than a single model independent of ϵ . In Proposition C.1: $\mathcal{M}_\epsilon$ has the same causal graph as $\mathcal{M}^*$; in the proof it is supplied by $\mathcal{M}_\delta$ with $\delta > 0$ small enough that $\delta/2 \leq \epsilon$. c denotes a locally specified scalar, categorical configuration, or real-valued function, according to its defining expression. In Proposition 4.1, Proposition C.1: $c > 0$ is chosen together with $\mathcal{M}^*$ before quantifying over $\epsilon > 0$, and is independent of ϵ ; it is a strict lower bound for $|\text{ATE}_{\mathcal{M}^*} - \text{ATE}_{\mathcal{M}_\epsilon}|$. In Lemma E.5: In the proof's final self-term calculation, $c = W_{1,\lambda}(\nu, \nu)$ is the entropy-regularized self-transport value, with $|c| \leq (1 + 2 \log m)\lambda$; it is nonpositive under the specified entropy convention. In Lemma E.6: $c > 0$ is the concentration constant multiplying $n\epsilon^2$ in the negative term of the probability exponent; it depends only on d, K and is independent of n, δ, ϵ . In Proposition 4.3: In the proof, $c(t) = \int \hat{f}_y(a, t) \nu_A(da) - \int \hat{f}_y(a, t) \mu_A(da)$ is the contribution from changing the observational marginal of A , with ν_A and μ_A the respective observational marginals. In Theorem 4.2:

In the observational concentration argument, c ranges over the fixed finite set of possible complete categorical label vectors: $p_c = P(C = c)$ and $\mu_c = P((U_1, \dots, U_{n_U}) \mid C = c)$, the latter being used for $p_c > 0$. In Corollary 4.4: In categorical types (u, c) and conditioning expressions, c is a complete categorical label vector, with $p_c = P(C = c)$ and $\mu_c = P((U_1, \dots, U_{n_U}) \mid C = c)$ for $p_c > 0$. Separately, in the concentration exponent $-cC^2m_nr_n^2$, $c > 0$ is a fixed concentration constant.

Here, $\mathcal{M}^*$ is the ground-truth model and $\mathcal{M}_\epsilon$ are the models we use to approximate $\mathcal{M}^*$. Proposition 4.1 implies that even if the observation distributions are close and the ground-truth model is identifiable, their ATEs can be far away, regardless of the sample size, which indicates that the method may not be consistent. We need some regularization on $\mathcal{M}_\epsilon$ to recover consistency. In particular, we find that regularizing the Lipschitz constant of the NN can help. Much work has been done to impose Lipschitz regularization during the training process [13, 49, 22, 42, 10]. We denote $\text{Lip}(f)$ to be the Lipschitz constant of a function f and define the truncated Lipschitz NN class,

$$\mathcal{NN}_{d_1, d_2}^{L_f, K}(W, L) = \{\max\{-K, \min\{f, K\}\} : f \in \mathcal{NN}_{d_1, d_2}(W, L), \text{Lip}(f) \leq L_f\}.$$

For simplicity, we omit the dimensions and use shorthand $\mathcal{NN}^{L_f, K}(W, L)$. Similarly, we also define the class $\mathcal{NN}^{\infty, K}(W_1, L_1, W_2, L_2, \tau)$ as a truncated version of $\mathcal{NN}(W_1, L_1, W_2, L_2, \tau)$. The next theorem gives the consistency result of the min estimator. To state the theorem, we define $F_* = \mathbb{E}_{t \sim \mu_T} \mathbb{E}_{\mathcal{M}^*}[F(V_1(t), \dots, V_{n_V}(t))]$ to be the true value and $\underline{F}^L$ to be the optimal value of the following optimization problem over SCMs with Lipschitz constant L .

$$\begin{aligned} \underline{F}^L = \underset{\widehat{\mathcal{M}} \in \mathcal{M}(\mathcal{G}, \mathcal{F}_L^K, \mathbf{U}), P(\mathbf{U})}{\text{argmin}} \quad & \mathbb{E}_{t \sim \mu_T} \mathbb{E}_{\widehat{\mathcal{M}}}[F(V_1(t), \dots, V_{n_V}(t))], \\ \text{s.t. } & W(P^{\widehat{\mathcal{M}}}(\mathbf{V}), P^{\mathcal{M}^*}(\mathbf{V})) = 0, \end{aligned} \quad (4.2)$$

where the minimum is taken over $\mathcal{M}(\mathcal{G}, \mathcal{F}_L^K, \mathbf{U})$ with $\mathcal{F}_L^K = \{f : \|f\| \leq K, f \in \mathcal{F}_L\}$ and all latent distributions $P(\mathbf{U})$. Note that if L is the Lipschitz bound L_f that we assume on our structural functions, then $\underline{F}^L$ is the sharp lower bound.

Theorem 4.2. *Let $\mathcal{M}^*$ be any SCM satisfying the assumptions of Corollary 3.8. Suppose that the Lipschitz constant of functions in $\mathcal{M}$ is L_f , $F : \mathbb{R}^{n_V} \rightarrow \mathbb{R}$ in (4.1) is Lipschitz continuous and $\tau_n > 0$, let $K > 0$ be the constant in ?? 3.2, $\widehat{L}_f = \sqrt{d_{\max}^{\text{in}} d_{\max}^{\text{out}}} L_f$,*

$$\mathcal{F}_{0,n} = \mathcal{NN}^{\widehat{L}_f, K}(W_{0,n}, \Theta(\log d_{\max}^{\text{in}})), \quad \mathcal{F}_{1,n} = \mathcal{NN}^{\infty, K}(\Theta(d_{\max}^U), L_{1,n}, \Theta((d_{\max}^U)^2), L_{2,n}, \tau_n),$$

where $d_{\max}^U$, $d_{\max}^{\text{in}}$ and $d_{\max}^{\text{out}}$ are defined in Corollary 3.8. Suppose in addition that $d_{\max}^{\text{in}} \geq 1$ and $d_{\max}^U \geq 1$. In (4.1), restrict each categorical raw structural network $q \in \mathcal{F}_{0,n}$ to satisfy $q(x) \in [0, \infty)^{n_i}$, $\|q(x)\|_1 > 0$ for every structural input x , and

$$\text{Lip} \left(x \mapsto \frac{q(x)}{\|q(x)\|_1} \right) \leq \hat{L}_f.$$

Put

$$D_U = \sum_{j=1}^{n_U} d_j^U, \quad D = \max\{1, D_U\},$$

where the sum defining D_U includes the non-Gumbel latent variables only. Take the radius to be $\alpha_n = \epsilon_n + s_n$, with positive tolerances chosen as sufficiently large constant multiples of the respective quantities in

$$\epsilon_n = O \left(W_{0,n}^{-1/d_{\max}^{\text{in}}} + \sum_{i=1}^2 L_{i,n}^{-2/d_{\max}^U} + \tau_n(1 - \log \tau_n) \right),$$

$$s_n = O \left(m_n^{-1/(D+2)} \log m_n + \frac{(L_{1,n} + L_{2,n}) \log m_n}{\sqrt{m_n}} + n^{-1/\max\{2, D_U\}} \log^2 n + \log(nm_n) \lambda_n \right).$$

For every continuous structural function f_i of $\mathcal{M}^*$, suppose in addition that $\sup_x \|f_i(x)\|_1 \leq K$ on its full structural input domain. Restrict every continuous structural network $\hat{f}_i \in \mathcal{F}_{0,n}$ in (4.1) to satisfy this same uniform output 1-norm bound. Retain the output bound in (4.2), namely

$$\mathcal{F}_L^K = \{f : \sup_x \|f(x)\|_1 \leq K, \text{Lip}(f) \leq L\}.$$

Let F_n be the optimal value of the empirical optimization with the objective and Sinkhorn constraint of (4.1) and both the continuous and categorical restrictions just specified, and let $\underline{F}^{\hat{L}_f}$ be the optimal value of (4.2) with this bounded Lipschitz class. If $m_n = \Omega(n)$, $L_{i,n} = \Theta(m_n^{d_{\max}^U/(2d_{\max}^U+4)})$, $i = 1, 2$, $\lim_{n \rightarrow \infty} \min\{\tau_n^{-1}, W_{0,n}, (\log(nm_n)\lambda_n)^{-1}\} = \infty$, then with probability 1, $[\liminf_{n \rightarrow \infty} F_n, \limsup_{n \rightarrow \infty} F_n] \subset [\underline{F}^{\hat{L}_f}, F_*]$.

Additional conditions. In this result the bound K applies to every SCM of the model class over which the optimization ranges, not only to $\mathcal{M}^*$: every variable has norm at most K , so each categorical V_i with labels in $[n_i]$ has $n_i \leq K$, and every intervention value lies in $[-K, K]$. The expectation defining F_* is a finite real iterated expectation. The inner objective is integrable under

$P(\mathbf{U})$ for μ_T -almost every intervention value, and the resulting real conditional-expectation function is integrable under μ_T . Almost-everywhere versions are allowed. Every objective expectation used in (4.1) is a finite real iterated expectation. For each candidate over which the objective ranges, the inner objective is integrable under its latent law for μ_T -almost every intervention value, and the resulting real conditional-expectation function is integrable under μ_T . Almost-everywhere versions are allowed. Every objective expectation used in (4.2) is a finite real iterated expectation. For each candidate over which the objective ranges, the inner objective is integrable under its latent law for μ_T -almost every intervention value, and the resulting real conditional-expectation function is integrable under μ_T . Almost-everywhere versions are allowed; one-sided infinite expectations are outside this real objective domain. The parameter objective $f(\theta) = \mathbb{E}_{t \sim \mu_T} \mathbb{E}_{\mathcal{M}^\theta} [F(\mathbf{V}(t))]$ is used on its finite real iterated-expectation domain: for every parameter on which this objective is used, the inner objective is integrable under the corresponding latent law for μ_T -almost every intervention value, and its real conditional-expectation function is integrable under μ_T . Almost-everywhere versions are allowed. This is the same expectation convention as for F_* , F_n , and $\underline{F}^L$. In the categorical structural equations, $\arg \max$ selects the smallest maximizing label when there is a tie. This is a measurable completion on ties; under the stipulated independent standard Gumbel noises and nonnegative propensity weights of positive total, it leaves every fixed-intervention law unchanged.

Remark. For each n , every candidate is evaluated on the same m_n draws of all its exogenous inputs, including its uniform inputs, generator Gumbel inputs, and categorical Gumbel inputs. The common exogenous draws are independent and identically distributed according to the specified product law of the exogenous inputs. The common simulation draws are independent of the observational sample. The observational empirical law $P_n^{\mathcal{M}^*}(\mathbf{V})$ is formed from n independent complete observations with common law $P^{\mathcal{M}^*}(\mathbf{V})$. The empirical optimization ranges over network representations, including distinct representations with the same latent population laws. F_n denotes the infimum of objective values, with an attained minimum having the same value. For a zero categorical propensity, $\log 0 = -\infty$. Ties in a categorical $\arg \max$ are resolved by the smallest label. The orders for ϵ_n and s_n are asymptotic as $n \rightarrow \infty$. The depth expression $\Theta(\log d_{\max}^{\text{in}})$ means $\Theta(1 + \log d_{\max}^{\text{in}})$ at $d_{\max}^{\text{in}} = 1$. The finite component counts and component cube dimensions of the latent generator architectures in $\mathcal{F}_{1,n}$ are fixed as n varies. The architectural Θ orders use sufficiently large fixed constant factors supplied by the network constructions in the proof.

The bound K applies throughout the optimization class to the observed variables V_i and indexed non-Gumbel latent variables U_j , with $n_i \leq K$ for categorical nodes and admissible intervention coordinates in $[-K, K]$; it does not bound the separately supplied standard Gumbel noises. $F_n = \inf \left\{ \mathbb{E}_{t \sim \mu_T} \mathbb{E}_{\widehat{\mathcal{M}}} [F(V_1(t), \dots, V_{n_V}(t))] : \widehat{\mathcal{M}} \in \text{NCM}_{\mathcal{G}}(\mathcal{F}_{0,n}, \mathcal{F}_{1,n}), S_{\lambda_n}(P_{m_n}^{\widehat{\mathcal{M}}}(\mathbf{V}), P_n^{\mathcal{M}^*}(\mathbf{V})) \leq \alpha_n \right\}$. Truncation means coordinatewise clipping of the final output to $[-K, K]$; in $\mathcal{NN}^{\infty, K}$ no finite Lipschitz bound is imposed. $\underline{F}^L$ denotes the constrained infimum of the displayed scalar objective over the stated canonical SCM class and its admissible latent laws, rather than a minimizing SCM. The coordinates $P(U_j)$ in $\theta_{\mathcal{M}}$ are the population laws of the non-Gumbel latent variables, not necessarily uniform laws; the independent standard Gumbel laws are fixed and omitted from this population tuple. The continuous structural functions and categorical propensity functions of $\mathcal{M}^*$ have Lipschitz constants at most L_f in the paper's 1-norm.

Additional conditions. For a result cited in the proof of Theorem 4.2: Every objective expectation used in (4.1) is a finite real iterated expectation. For each candidate over which the objective ranges, the inner objective is integrable under its latent law for μ_T -almost every intervention value, and the resulting real conditional-expectation function is integrable under μ_T . Almost-everywhere versions are allowed. For a result cited in the proof of Theorem 4.2: The observational empirical law $P_n^{\mathcal{M}^*}(\mathbf{V})$ is formed from n independent complete observations with common law $P^{\mathcal{M}^*}(\mathbf{V})$.

Remark (notation). ϵ_n denotes a positive deterministic approximation-tolerance sequence indexed by $n \in \mathbb{N}$, with its approximation role specified locally. In Proposition C.2: $\epsilon_n > 0$ for $n \in \mathbb{N}$, $\epsilon_n \rightarrow 0$, and it bounds the distance from each point of $\widetilde{\Theta}_{\infty}$ to an available point of Θ_n . In Theorem 4.2, Corollary 4.4: In the radius definition, $\alpha_n = \epsilon_n + s_n$, where $\epsilon_n > 0$ is chosen with sufficiently large constants in $\epsilon_n = O\left(W_{0,n}^{-1/d_{\max}^n} + \sum_{i=1}^2 L_{i,n}^{-2/d_{\max}^U} + \tau_n(1 - \log \tau_n)\right)$. In a result cited in the proof of Theorem 4.2: In the empirical generator concentration argument, ϵ_n is a local deterministic deviation tolerance in the bound $\sup_{\theta_{i,j}} W(P(f_{\theta_{i,j}}(Z_{i,j})), P_{m_n}(f_{\theta_{i,j}}(Z_{i,j}))) \leq \epsilon_n + \epsilon'_n$; it is chosen as $\epsilon_n = C m_n^{-1/(d_{\max}^U+2)} \log m_n$ with fixed $C > 0$ sufficiently large. This local choice is distinct from the architecture approximation tolerance in the constraint radius. s_n is a deterministic tolerance sequence indexed by n , specified by its displayed asymptotic order rather than by a unique numerical formula. In Theorem 4.2: $s_n > 0$ is chosen as a sufficiently large constant multiple of the quantity in the displayed order $O\left(m_n^{-1/(D+2)} \log m_n + \frac{(L_{1,n}+L_{2,n}) \log m_n}{\sqrt{m_n}} + n^{-1/\max\{2, D_U\}} \log^2 n + \log(n m_n) \lambda_n\right)$, where $D_U = \sum_{j=1}^{n_U} d_j^U$ includes the non-Gumbel latent variables only and $D = \max\{1, D_U\}$. In Corollary 4.4: $s_n > 0$ is chosen as a sufficiently large constant multiple of the quantity in the

displayed order $O\left(m_n^{-1/(D+2)} \log m_n + \frac{(L_{1,n}+L_{2,n}) \log m_n}{\sqrt{m_n}} + n^{-1/\max\{2, D_U\}} \log^2 n + \log(nm_n)\lambda_n\right)$, where $D_U = \sum_{j=1}^{n_U} d_j^U$ includes the non-Gumbel latent variables only and $D = \max\{1, D_U\}$. τ_n is a sequence indexed by n , used locally either as a generator temperature or as an additive tolerance in an approximate Lipschitz bound. In Proposition C.2: τ_n is the additive tolerance in equation C.1, and $\tau_n \rightarrow 0$. In Theorem 4.2, Corollary 4.4: In the network architecture and growth conditions, $\tau_n > 0$ is the Gumbel-softmax generator temperature and $\tau_n \rightarrow 0$; in the proof's application of the optimization proposition, its additive tolerance is assigned the separate value $4(1 + \log(nm_n))\lambda_n$. In a result cited in the proof of Theorem 4.2: When verifying equation C.1, τ_n denotes the additive tolerance $4(1 + \log(nm_n))\lambda_n$, which tends to zero under the stated regularization condition; this assignment does not identify it with the generator temperature. $W_{0,n}$ is the width parameter of the structural-function approximation class $\mathcal{F}_{0,n}$, indexed by sample size n . In Theorem 4.2, Corollary 4.4: $W_{0,n} \rightarrow \infty$ as $n \rightarrow \infty$, as required by $\lim_{n \rightarrow \infty} \min\{\tau_n^{-1}, W_{0,n}, (\log(nm_n)\lambda_n)^{-1}\} = \infty$. $L_{i,n}$, $i = 1, 2$, indexes the depth sequences of the first and second parts of the latent-distribution generator architecture, with n the sample-size index. In Theorem 4.2, Corollary 4.4: $L_{i,n} = \Theta(m_n^{d_{\max}^U/(2d_{\max}^U+4)})$ for $i = 1, 2$, with $m_n = \Omega(n)$. s denotes the fixed integer depth parameter $s = 1$, distinct from the quantization parameter δ and the quantization mesh δ/n .

Note that our theorem shows that the lower limit of the solution is large than the point $\underline{F}^{\hat{L}_f}$, where $\hat{L}_f$ is slightly larger than the original constraint L_f we impose on the structural functions. Hence, this point can potentially be slightly smaller than the sharp bound $\underline{F}^{L_f}$. This worsening is due to the fact that we need to use NNs that satisfy a slightly worse Lipschitz property, to ensure that we have sufficient approximation power. Although Theorem 4.2 does not guarantee $\{F_n\}$ converges to a point, it states that $\{F_n\}$ may oscillate in the interval $[\underline{F}^{\hat{L}_f}, F_*]$, which will still give a useful lower bound to ground-truth value F_* . In particular, if the graph is identifiable, we have $F^* = \underline{F}^{\hat{L}_f}$ and $\{F_n\}$ converges to F^* . Also, note that in Theorem 4.2, we use wide NNs rather than deep NN for $\mathcal{F}_{0,n}$ because results in [52] show that wide NNs can approximate Lipschitz functions while controlling their Lipschitz constants (a property that is not yet established for deep NNs). Similar results can be obtained if wide neural nets are used for all components, invoking Theorem 3.6.

Remark (notation). $\mathcal{F}_L^K(K_1, K_2)$ denotes a class of bounded Lipschitz continuous maps $f : K_1 \rightarrow K_2$, with the norm bound and Lipschitz restriction given by its local definition. In Theorem 4.2: In the statement and the limiting optimization, $\mathcal{F}_L^K = \{f : \sup_x \|f(x)\|_1 \leq K, \text{Lip}(f) \leq L\}$, with the supremum over the full structural input domain; in the proof's displayed population-function

definition, $\mathcal{F}_L^K = \{f : \|f\|_\infty \leq K, f \text{ Lipschitz continuous}\}$, where the function infinity norm is the uniform coordinatewise bound and that display supplies no numerical Lipschitz bound. In Corollary 4.4: The population bounded Lipschitz class is $\mathcal{F}_L^K = \{f : \sup_x \|f(x)\|_1 \leq K, \text{Lip}(f) \leq L\}$, with the supremum over the full structural input domain. $\mathcal{NN}^{\infty,K}$ clips the final generator output coordinatewise to $[-K, K]$, with no imposed Lipschitz bound; its suppressed dimensions and component structure are those of the corresponding generator class. F_n is the scalar infimum of $\mathbb{E}_{t \sim \mu_T} \mathbb{E}_{\widehat{\mathcal{M}}} [F(V_1(t), \dots, V_{n_V}(t))]$ over the candidate neural causal models admitted by the empirical optimization, subject to $S_{\lambda_n}(P_{m_n}^{\widehat{\mathcal{M}}}(\mathbf{V}), P_n^{\mathcal{M}^*}(\mathbf{V})) \leq \alpha_n$, wherever this value is defined; an attained minimum has the same value. In Corollary 4.4: F_n is the optimal value of the locally specified restricted optimization, retaining the objective and Sinkhorn constraint of (4.1) and using the specified classes $\mathcal{F}_{0,n}$ and $\mathcal{F}_{1,n}$. The minimized objective is $\frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \mathbb{E}_{\widehat{\mathcal{M}}} [Y(t)] dt$. Each continuous structural network $\widehat{f}_i \in \mathcal{F}_{0,n}$ satisfies $\sup_x \|\widehat{f}_i(x)\|_1 \leq K$ on its full structural input domain. Each categorical raw structural network $q \in \mathcal{F}_{0,n}$ satisfies $q(x) \in [0, \infty)^{n_i}$, $\|q(x)\|_1 > 0$ for every structural input x , and $\text{Lip}(x \mapsto q(x)/\|q(x)\|_1) \leq \widehat{L}_f$. For each sample size, every candidate empirical law is obtained by evaluating the candidate on the same m_n independent draws of its uniform and Gumbel exogenous inputs, independent of the observational sample. In Theorem 4.2: The optimization ranges over network representations using $\mathcal{F}_{0,n}$ and $\mathcal{F}_{1,n}$. Each continuous structural network $\widehat{f}_i$ satisfies $\sup_x \|\widehat{f}_i(x)\|_1 \leq K$ on its full structural input domain. Each categorical raw structural network $q \in \mathcal{F}_{0,n}$ satisfies $q(x) \in [0, \infty)^{n_i}$, $\|q(x)\|_1 > 0$ for every structural input x , and $\text{Lip}(x \mapsto q(x)/\|q(x)\|_1) \leq \widehat{L}_f$. For each n , all candidates are evaluated on the same m_n independent and identically distributed draws of all exogenous inputs according to their specified product law, including uniform inputs, generator Gumbel inputs, and categorical Gumbel inputs, independent of the observational sample; representations with the same latent population laws may yield different empirical constraints. μ_T is the given measure on treatment intervention values used in the outer expectation $\mathbb{E}_{t \sim \mu_T}$. In Corollary 4.4: $\mu_T = \text{Unif}([t_1, t_2])$, $\mu_T \ll P$, and $A_T^2 = \int_{t_1}^{t_2} (d\mu_T/dP(t))^2 P(dt) < \infty$, where P is the reference probability measure on the treatment space. $F_* = \mathbb{E}_{t \sim \mu_T} \mathbb{E}_{\mathcal{M}^*} [F(V_1(t), \dots, V_{n_V}(t))]$ denotes the population objective evaluated at the true SCM. In Corollary 4.4: $F_* = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \mathbb{E}_{\mathcal{M}^*} [Y(t)] dt$, since $F(\mathbf{V}) = Y$ and $\mu_T = \text{Unif}([t_1, t_2])$. $\underline{F}^L$ is the population optimal objective value in equation 4.2, minimizing $\mathbb{E}_{t \sim \mu_T} \mathbb{E}_{\widehat{\mathcal{M}}} [F(V_1(t), \dots, V_{n_V}(t))]$ over canonical SCMs on $\mathcal{G}$ with structural functions in $\mathcal{F}_L^K$ and all latent distributions $P(\mathbf{U})$, subject to $W(P_{m_n}^{\widehat{\mathcal{M}}}(\mathbf{V}), P_n^{\mathcal{M}^*}(\mathbf{V})) = 0$. In Theorem 4.2: $\underline{F}^L$ uses the class $\mathcal{F}_L^K = \{f : \sup_x \|f(x)\|_1 \leq K, \text{Lip}(f) \leq L\}$ in equation 4.2, with the supremum taken over

the full structural input domain; categorical structural functions are normalized propensity vectors. The conclusion evaluates this population optimal value at $L = \widehat{L}_f$. $P_n^{\mathcal{M}^*}(\mathbf{V}) = n^{-1} \sum_{a=1}^n \delta_{\mathbf{V}^{(a)}}$ is the empirical observational law of the n observed joint values from $P^{\mathcal{M}^*}(\mathbf{V})$. In Theorem 4.2, Corollary 4.4: The n observations defining $P_n^{\mathcal{M}^*}(\mathbf{V})$ are independent complete observations with common law $P^{\mathcal{M}^*}(\mathbf{V})$. $P_{m_n}^{\widehat{\mathcal{M}}}(\mathbf{V})$ denotes the empirical observational law for the candidate $\widehat{\mathcal{M}}$, namely $m_n^{-1} \sum_{r=1}^{m_n} \delta_{\mathbf{V}^{(r)}}$, and is indexed by the candidate rather than denoting only the law at an optimizer. In Theorem 4.2: For each n , $P_{m_n}^{\widehat{\mathcal{M}}}(\mathbf{V})$ is obtained by evaluating every candidate network representation on the same m_n independent draws of all exogenous inputs, including uniform inputs, generator Gumbel inputs, and categorical Gumbel inputs; these draws are independent of the observational sample. Different network representations with the same latent population laws may give different common-draw empirical distributions. In Corollary 4.4: For each n , $P_{m_n}^{\widehat{\mathcal{M}}}(\mathbf{V})$ is obtained by evaluating every candidate network representation on the same m_n independent draws of all exogenous inputs, including uniform inputs, generator Gumbel inputs, and categorical Gumbel inputs; these draws are independent of the observational sample. Different network representations with the same latent population laws may give different common-draw empirical distributions. $\mathcal{F}_{0,n}$ is the prescribed structural network class at sample size n , with the input and output dimensions appropriate to the node and with truncation applied coordinatewise. In Theorem 4.2: In the parameter-space construction, a continuous-node candidate is a function $\widehat{f}_i \in \mathcal{F}_{0,n}$ satisfying $\sup_x \|\widehat{f}_i(x)\|_1 \leq K$ on its full structural input domain. A categorical-node propensity is represented by $x \mapsto q(x)/\|q(x)\|_1$ with $q \in \mathcal{F}_{0,n}$ restricted to satisfy $q(x) \in [0, \infty)^{n_i}$ and $\|q(x)\|_1 > 0$ for every structural input x , and $\text{Lip}(x \mapsto q(x)/\|q(x)\|_1) \leq \widehat{L}_f$. The depth expression $\Theta(\log d_{\max}^{\text{in}})$ uses $\Theta(1 + \log d_{\max}^{\text{in}})$ when $d_{\max}^{\text{in}} = 1$. In Corollary 4.4: A continuous-node candidate is a function $\widehat{f}_i \in \mathcal{F}_{0,n}$ satisfying $\sup_x \|\widehat{f}_i(x)\|_1 \leq K$ on its full structural input domain. A categorical-node propensity is represented by $x \mapsto q(x)/\|q(x)\|_1$ with $q \in \mathcal{F}_{0,n}$ restricted to satisfy $q(x) \in [0, \infty)^{n_i}$ and $\|q(x)\|_1 > 0$ for every structural input x , and $\text{Lip}(x \mapsto q(x)/\|q(x)\|_1) \leq \widehat{L}_f$. The depth expression $\Theta(\log d_{\max}^{\text{in}})$ uses $\Theta(1 + \log d_{\max}^{\text{in}})$ when $d_{\max}^{\text{in}} = 1$. $\mathcal{F}_{1,n}$ is the latent-generator network class indexed by n , with architecture $\mathcal{NN}^{\infty,K}(\Theta(d_{\max}^U), L_{1,n}, \Theta((d_{\max}^U)^2), L_{2,n}, \tau_n)$ and coordinatewise output truncation to $[-K, K]$. In Theorem 4.2, Corollary 4.4: The finite component counts and component cube dimensions of the latent generator architectures are fixed as n varies. $\lambda_n > 0$, indexed by the observational sample size n , is the regularization parameter in $S_{\lambda_n}(P_{m_n}^{\widehat{\mathcal{M}}}(\mathbf{V}), P_n^{\mathcal{M}^*}(\mathbf{V}))$. In Theorem 4.2, Corollary 4.4: $\log(nm_n)\lambda_n \rightarrow 0$ as $n \rightarrow \infty$. $\mathcal{N}$ denotes either the covering-number functional when supplied with covering arguments, or a locally quantified causal model. In

Lemma E.6, a result cited in the proof of Lemma E.6: $\mathcal{N}(\epsilon, S, \|\cdot\|_1)$ denotes the covering number of the evaluation set S in the summed 1-norm; the separately marked averaged covering numbers use the averaged empirical metric. In Proposition 4.3: In the proof, $\mathcal{N} \in \{\mathcal{M}, \widehat{\mathcal{M}}\}$ ranges over the two causal models when bounding their structural regression functions. In a result cited in the proof of Lemma E.6: $\mathcal{N}_1^{\text{avg}}(r, \mathcal{G}, N)$ is the supremum over $x_1, \dots, x_N \in X$ of the covering number of $\{(g(x_1), \dots, g(x_N)) : g \in \mathcal{G}\}$ in $d_1(u, w) = N^{-1} \sum_{i=1}^N |u_i - w_i|$; the concentration bound uses $r = \epsilon/16$ and $N = 2m$.

As a special case, we leverage Theorem 4.2 for a non-asymptotic rate for the ATE without confounding. The following proposition proves the Hölder continuity of ATE when there is no confounding.

Proposition 4.3 (Hölder continuity of ATE). *Given two causal models $\mathcal{M}, \widehat{\mathcal{M}} \in \mathcal{M}(\mathcal{G}, \mathcal{F}_L, U)$ satisfying ?? 2.3 and ?? 3.2, let their observational distributions be ν, μ . Suppose the norms of all variables are bounded by $K > 0$. If (1) (Overlap) ν is absolutely continuous with respect to one probability measure P and the density $p_\nu(t | \text{Pa}(T) = x) \geq \delta > 0$ for x almost surely and $t \in [t_1, t_2]$ and (2) (No Confounding) there is no hidden confounder in the causal graph $\mathcal{G}$, we have*

$$\int_{t_1}^{t_2} (\mathbb{E}_{\mathcal{M}}[Y(t)] - \mathbb{E}_{\widehat{\mathcal{M}}}[\widehat{Y}(t)])^2 P(dt) \leq \frac{2C_W}{\delta} W(\mu, \nu) + 2L_y^2 W^2(\mu, \nu) P([t_1, t_2]),$$

where

$$D = \max \left(\{1\} \cup \left\{ \frac{n_i - 1}{2} : V_i \text{ is categorical} \right\} \right), \quad L_y = (LD + 1)^{n_V},$$

$$q = \sum_{i: V_i \text{ is continuous}} d_i + |\{i : V_i \text{ is categorical}\}|, \quad B = K(1 + 2qL_y),$$

$$\text{and } C_W = 4L_y B + 2B \max\{L_y, 1\}.$$

Remark. The measure P is the reference probability measure on the treatment space: for $X = \text{Pa}(T)$, the conditional density notation means $\nu(dt | X = x) = p_\nu(t | X = x)P(dt)$. For each of $\mathcal{M}$ and $\widehat{\mathcal{M}}$, $\max_{i \in [n_V], j \in [n_U]} \{\|V_i\|_\infty, \|U_j\|_\infty\} \leq K$ almost surely, where the indexed U_j are the non-Gumbel latent variables. The measure P is a probability measure on the treatment space, and for almost every parent value x , $\nu(dt | \text{Pa}(T) = x) = p_\nu(t | \text{Pa}(T) = x)P(dt)$, with the stated density version satisfying $p_\nu(t | \text{Pa}(T) = x) \geq \delta$ for every $t \in [t_1, t_2]$.

Remark (notation). $\widehat{Y}$ denotes the real-valued outcome in the comparison distribution or comparison causal model. In Lemma C.3: $(\widehat{T}, \widehat{Y}) \sim \mu$, with $\widehat{f}(t) = \mathbb{E}_\mu[\widehat{Y} \mid \widehat{T} = t]$; the proof treats this pair as supported in $\{(t, y) : \|(t, y)\|_\infty \leq K\}$. In Proposition 4.3: $\widehat{Y}$ is the outcome coordinate generated by $\widehat{\mathcal{M}}$, whose observational law is μ ; $\widehat{Y}(t)$ denotes its outcome under intervention $T = t$, and $|\widehat{Y}| \leq K$. C_W is a numerical comparison coefficient whose defining formula depends on the context. In Lemma C.3: $C_W = 4LK + 2K \max\{L, 1\}$, where L bounds the Lipschitz constants of f and $\widehat{f}$ and K bounds their absolute values. In Proposition 4.3: $C_W = 4L_y B + 2B \max\{L_y, 1\}$, where $L_y = (LD + 1)^{n_V}$, $D = \max\left(\{1\} \cup \left\{\frac{n_i - 1}{2} : V_i \text{ is categorical}\right\}\right)$, $q = \sum_{i: V_i \text{ is continuous}} d_i + |\{i : V_i \text{ is categorical}\}|$, and $B = K(1 + 2qL_y)$. Writing $X = \text{Pa}(T)$, $p_\nu(t \mid X = x)$ is the conditional treatment density of $\nu(dt \mid x)$ with respect to the comparison measure $P(dt)$, so that $\nu(dt \mid x) = p_\nu(t \mid x)P(dt)$; its overlap lower bound holds for ν_X -almost every x and $t \in [t_1, t_2]$. $\delta > 0$ denotes a locally specified positive real parameter. In Proposition 4.3, Corollary 4.4: δ in the overlap premise is the lower bound satisfying $p_\nu(t \mid \text{Pa}(T) = x) \geq \delta$ for x almost surely and $t \in [t_1, t_2]$. In Lemma E.4: δ is the auxiliary threshold for the gaps between $G_i + \log p_i$; at the end of the proof it is chosen as $\delta = -\tau \log \tau$. In Proposition C.1: In the proof, δ parameterizes $\mathcal{M}_\delta$: the covariate has an atom at δ , and the outcome equation contains X/δ for $X \geq 0$; the construction takes $\delta \rightarrow 0$. In Lemma E.6: δ is an arbitrary positive tolerance; the probability bound is for each fixed pair $\delta, \epsilon > 0$. In the proof, $\mathcal{H}_\delta$ has uniform-norm covering radius $\delta/2$, so replacing a test function by its representative contributes at most δ to the difference of its empirical and population integrals. In Theorem 4.2: In the simulation-concentration argument, the empirical Wasserstein bound is applied with $\delta = r_n$, where $H_n = 1 + L_{1,n} + L_{2,n}$ and $r_n = m_n^{-1/(D+2)} \log m_n + H_n m_n^{-1/2} \log m_n$; its other tolerance is $\epsilon = A_0 r_n$. In a result cited in the proof of Theorem 3.6: For a component of positive dimension $q = d_i^C$, $\delta = 1/\lceil \sqrt{q}(q+1)n(n-1) \rceil$ is the source quantization parameter, and the quantization mesh is δ/n ; here $n = \lfloor (W_1/A_q)^{1/q} \rfloor$ for $W_1 \geq A_q 2^q$, with $A_q \geq 1$ depending only on q . In Corollary 4.4: Separately from the overlap parameter, the proof's simulation-concentration argument chooses the local tolerance $\delta = r_n$, where $H_n = 1 + L_{1,n} + L_{2,n}$ and $r_n = m_n^{-1/(D+2)} \log m_n + H_n m_n^{-1/2} \log m_n$, for sufficiently large n ; the other tolerance is $\epsilon = C r_n$ for a sufficiently large fixed positive constant C . t_1 is a real scalar whose role is local to the intervention interval or the Lipschitz comparison. In Proposition 4.3, Corollary 4.4: t_1 is the fixed lower endpoint of the intervention interval $[t_1, t_2]$. In Corollary 4.4: $t_1 < t_2$, and μ_T is the uniform probability law on $[t_1, t_2]$. In Lemma C.3: In the proof, t_1 is the treatment coordinate of the first arbitrary comparison point (t_1, y_1) in $\{(t, y) : \|(t, y)\|_\infty \leq K\}$. t_2 is a real scalar whose role is

local to the intervention interval or the Lipschitz comparison. In Proposition 4.3, Corollary 4.4: t_2 is the fixed upper endpoint of the intervention interval $[t_1, t_2]$. In Corollary 4.4: $t_1 < t_2$, and μ_T is the uniform probability law on $[t_1, t_2]$. In Lemma C.3: In the proof, t_2 is the treatment coordinate of the second arbitrary comparison point (t_2, y_2) in $\{(t, y) : \|(t, y)\|_\infty \leq K\}$.

As a corollary of Proposition 4.3 and Theorem 4.2, we can give a convergence rate for the class of estimands in Theorem 4.2.

Corollary 4.4. *Let F, μ_T in Theorem 4.2 be $F(\mathbf{V}) = Y$, $\mu_T \sim \text{Unif}([t_1, t_2])$, and $\epsilon > 0$. Suppose that $t_1 < t_2$. Suppose that the assumptions in Proposition 4.3 and Theorem 4.2 hold. Suppose in addition that $\mu_T \ll P$ and*

$$A_T^2 = \int_{t_1}^{t_2} \left(\frac{d\mu_T}{dP}(t) \right)^2 P(dt) < \infty,$$

where P is the reference probability measure in Proposition 4.3. Use the network classes $\mathcal{F}_{0,n}$ and $\mathcal{F}_{1,n}$ and the population structural classes specified in Theorem 4.2. In (4.1), restrict each continuous structural network $\hat{f}_i \in \mathcal{F}_{0,n}$ to satisfy $\sup_x \|\hat{f}_i(x)\|_1 \leq K$ on its full structural input domain, and restrict each categorical raw structural network $q \in \mathcal{F}_{0,n}$ to satisfy $q(x) \in [0, \infty)^{n_i}$, $\|q(x)\|_1 > 0$ for every structural input x , and

$$\text{Lip} \left(x \mapsto \frac{q(x)}{\|q(x)\|_1} \right) \leq \hat{L}_f.$$

For each sample size, obtain every candidate empirical law by evaluating the candidate on the same m_n independent draws of its uniform and Gumbel exogenous inputs, independent of the observational sample. Let F_n be the optimal value of this restricted optimization, retaining the objective and Sinkhorn constraint of (4.1). Put

$$D_U = \sum_{j=1}^{n_U} d_j^U, \quad D = \max\{1, D_U\},$$

where the sum includes only the non-Gumbel latent variables. Take $\alpha_n = \epsilon_n + s_n$, with positive tolerances chosen as sufficiently large constant multiples of the respective quantities in

$$\epsilon_n = O \left(W_{0,n}^{-1/d_{\max}^{\text{in}}} + \sum_{i=1}^2 L_{i,n}^{-2/d_{\max}^U} + \tau_n(1 - \log \tau_n) \right),$$

$$s_n = O \left(m_n^{-1/(D+2)} \log m_n + \frac{(L_{1,n} + L_{2,n}) \log m_n}{\sqrt{m_n}} + n^{-1/\max\{2, D_U\}} \log^2 n + \log(nm_n) \lambda_n \right).$$

Then, with probability at least $1 - O(n^{-2})$, we have $|F_n - F_*| \leq O(\sqrt{\alpha_n})$, where F_* is defined in Theorem 4.2 and F_n, α_n are specified here.

Additional conditions. In this result the bound K applies to every SCM of the model class over which the optimization ranges, not only to $\mathcal{M}^*$: every variable has norm at most K , so each categorical V_i with labels in $[n_i]$ has $n_i \leq K$, and every intervention value lies in $[-K, K]$. The expectation defining F_* is a finite real iterated expectation. The inner objective is integrable under $P(\mathbf{U})$ for μ_T -almost every intervention value, and the resulting real conditional-expectation function is integrable under μ_T . Almost-everywhere versions are allowed. Every objective expectation used in (4.1) is a finite real iterated expectation. For each candidate over which the objective ranges, the inner objective is integrable under its latent law for μ_T -almost every intervention value, and the resulting real conditional-expectation function is integrable under μ_T . Almost-everywhere versions are allowed. Every objective expectation used in (4.2) is a finite real iterated expectation. For each candidate over which the objective ranges, the inner objective is integrable under its latent law for μ_T -almost every intervention value, and the resulting real conditional-expectation function is integrable under μ_T . Almost-everywhere versions are allowed; one-sided infinite expectations are outside this real objective domain. In the categorical structural equations, $\arg \max$ selects the smallest maximizing label when there is a tie. This is a measurable completion on ties; under the stipulated independent standard Gumbel noises and nonnegative propensity weights of positive total, it leaves every fixed-intervention law unchanged.

Remark. The observational empirical law $P_n^{\mathcal{M}^*}(\mathbf{V})$ is formed from n independent complete observations with common law $P^{\mathcal{M}^*}(\mathbf{V})$. The variable bound inherited from Assumption 3.2 uses $\|\cdot\|_\infty$. $\mu_T = \text{Unif}([t_1, t_2])$ is the deterministic intervention probability measure.

5 Experiments

In this section, we examine the performance of our algorithm in two settings. We compare our algorithm with the Autobounds algorithm [14] in a binary IV example in [14] and in a continuous IV model to calculate bounds for ATE¹. Since Autobounds can only deal with discrete models, we

¹The code can be found in <https://github.com/Jiyuan-Tan/NeuralPartialID>

discretize the continuous model for comparison purposes. The implementation details are provided in the Section D. We also provide an extra experiment on the counterexample of Proposition C.1 in the appendix to show the effect of Lipschitz regularization.

5.1 Discrete IV

The setting of the first experiment is taken from [14, Section D.1]. This is a binary IV problem and we can calculate the optimal bound using LP [4]. The causal graph is shown in Figure 3. We find that our bound is close to the optimal bound. Both algorithms cover the true value in all runs. We could see from Table 1 that the bound given by NCM is slightly worse than Autobounds, but is still close to the optimal bound. Besides, the average Autobounds bound does not cover the optimal bound in this example.

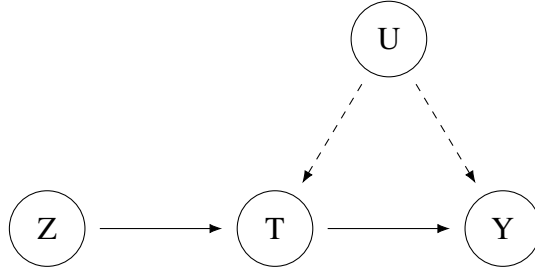


Figure 3: Instrumental Variable (IV) graph. Z is the instrumental variable, T is the treatment and Y is the outcome.

Algorithm	Average Bound	SD of length	Optimal Bound	True Value
NCM (Ours)	$[-0.49, 0.05]$	0.05	$[-0.45, -0.04]$	-0.31
Autobounds ([14])	$[-0.45, -0.05]$	0.02	$[-0.45, -0.04]$	-0.31

Table 1: The sample size is taken to be 5000. We average over 10 runs with different random seeds. SD is short for the standard deviation. In all runs, the bounds given by the two algorithms all cover the true value.

5.2 Continuous IV

Now, we turn to the continuous IV setting, where T is still binary, but Y and Z can be continuous. Let $E_i^\lambda \sim \lambda Z_1 + (1 - \lambda)\text{Unif}(-1, 1)$, where Z_1 is the Gaussian variable conditioning on $[-1, 1]$ and the structure equations of $\mathcal{M}^\lambda$ to be

$$\begin{aligned} E_Y &= E_1^\lambda, U = E_2^\lambda, Z = E_3^\lambda, \\ P(T = 1|Z) &= (1.5 + Z + 0.5U)/3, \\ Y &= 3T - 1.5TU + U + E_Y, \end{aligned}$$

where E_i^λ are independent. It is easy to see the ATE of this model is 3 regardless of λ . In the experiment, we randomly choose ten $\lambda \sim \text{Unif}(0, 1)$. For each λ , we run the algorithms 5 times to get the bound. We choose different latent distributions (indexed by λ) in the experiments to create more difficulty.

Since Autobounds can only deal with discrete variables, we discretize Z and Y . Suppose that $Z \in [l, u]$, we map all points in intervals $[l + i(u - l)/k, l + (i + 1)(u - l)/k], I = 0, 1, \dots, k - 1$ to the middle point $l + (i + 1/2)(u - l)/k$. We choose $k = 8$ in the following experiments. The choice will give rise to polynomial programming problems with 2^{14} decision variables, which is quite large.

Table 2 demonstrates the results. While both algorithms cover the true ATE well, we can see that NCM gives much tighter bounds on average. The main reason may be that the discretized problem does not approximate the original problem well enough. It is possible that a larger discretized parameter k can give a better bound, but since the size of the polynomial problem grows exponentially with k , the optimization problem may be intractable for large k . On the contrary, NCM does not suffer from computational difficulties.

6 Conclusion

In this paper, we provide theoretical justification for using NCMs for partial identification. We show that NCMs can be used to represent SCMs with complex unknown latent distributions under mild assumptions and prove the asymptotic consistency of the max/min estimator for partial identification of causal effects in general settings with both discrete and continuous variables. Our results also provide guidelines on the practical implementation of this method and on what hyperparameters are

Algorithm	Average Bound	SD of length	Success Rate	True Value
NCM (Ours)	[2.49,3.24]	0.49	50/50	3
Autobounds ([14])	[1.40, 3.48]	0.26	50/50	3

Table 2: We take a sample size of 5000. We randomly choose 10 $\lambda \sim \text{Unif}(0, 1)$ and get 10 models $\mathcal{M}^{\lambda_i}$. For each model $\mathcal{M}^{\lambda_i}$, we run the two algorithms 5 times. The success rate is the number of times when the obtained bounds cover the true ATE divided by the total number of experiments. SD is short for the standard deviation.

important, as well as recommendations on values that these hyperparameters should take for the consistency of the method. These practical guidelines were validated with a small set of targeted experiments, which also showcase superior performance of the neural-causal approach as compared to a prior main contender approach from econometrics and statistics, that involves discretization and polynomial programming.

An obvious next step in the theoretical foundation of neural-causal models is providing finite sample guarantees for this method, which requires substantial further theoretical developments in the understanding of the geometry of the optimization program that defines the bounds on the causal effect of interest. We take a first step in that direction for the special case, when there are no unobserved confounders and view the general case as an exciting avenue for future work.

Acknowledgement Vasilis Syrgkanis is supported by NSF Award IIS-2337916 and a 2022 Amazon Research Award.

References

- [1] J. Angrist and G. Imbens. Identification and estimation of local average treatment effects, 1995.
- [2] M. Anthony, P. L. Bartlett, P. L. Bartlett, et al. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.
- [3] V. Balazadeh, V. Syrgkanis, and R. G. Krishnan. Partial identification of treatment effects with implicit generative models, Oct. 2022.

- [4] A. Balke and J. Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty Proceedings 1994*, pages 46–54. Elsevier, 1994.
- [5] A. Balke and J. Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, Sept. 1997.
- [6] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.
- [7] A. Bellot. Towards bounding causal effects under markov equivalence. *arXiv preprint arXiv:2311.07259*, 2023.
- [8] B. Bonet. Instrumentality tests revisited, Jan. 2013.
- [9] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.
- [10] L. Bungert, R. Raab, T. Roith, L. Schwinn, and D. Tenbrinck. CLIP: Cheap Lipschitz Training of Neural Networks. In A. Elmoataz, J. Fadili, Y. Quéau, J. Rabin, and L. Simon, editors, *Scale Space and Variational Methods in Computer Vision*, volume 12679, pages 307–319, Cham, 2021. Springer International Publishing.
- [11] V. Chernozhukov, H. Hong, and E. Tamer. Estimation and confidence regions for parameter sets in econometric models 1. *Econometrica*, 75(5):1243–1284, 2007.
- [12] L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269, 2020.
- [13] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier. Parseval networks: Improving robustness to adversarial examples. In *International conference on machine learning*, pages 854–863. PMLR, 2017.
- [14] G. Duarte, N. Finkelstein, D. Knox, J. Mummolo, and I. Shpitser. An automated approach to causal inference in discrete settings, Sept. 2021.

- [15] J. Feydy. Geometric data analysis, beyond convolutions. *Applied Mathematics*, 2020.
- [16] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trounev, and G. Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.
- [17] N. Fournier and A. Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3-4):707–738, 2015.
- [18] D. Frauen, T. Hatt, V. Melnychuk, and S. Feuerriegel. Estimating average causal effects from patient trajectories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7586–7594, 2023.
- [19] D. Frauen, F. Imrie, A. Curth, V. Melnychuk, S. Feuerriegel, and M. van der Schaar. A Neural Framework for Generalized Causal Sensitivity Analysis, Apr. 2024.
- [20] D. A. Freedman. *Statistical models and causal inference: a dialogue with the social sciences*. Cambridge University Press, 2010.
- [21] L.-A. Gottlieb, A. Kontorovich, and R. Krauthgamer. Efficient regression in metric spaces via approximate lipschitz extension. *IEEE Transactions on Information Theory*, 63(8):4838–4849, 2017.
- [22] H. Gouk, E. Frank, B. Pfahringer, and M. J. Cree. Regularisation of neural networks by enforcing Lipschitz continuity. *Machine Learning*, 110(2):393–416, Feb. 2021.
- [23] H. Gouk, E. Frank, B. Pfahringer, and M. J. Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110:393–416, 2021.
- [24] F. Gunsilius. A path-sampling method to partially identify causal effects in instrumental variable models, June 2020.
- [25] C. H. Hamilton and A. Rau-Chaplin. Compact hilbert indices for multi-dimensional data. In *First International Conference on Complex, Intelligent and Software Intensive Systems (CISIS’07)*, pages 139–146. IEEE.
- [26] S. Han and S. Yang. A computational approach to identification of treatment effects for policy evaluation. *Journal of Econometrics*, 240(1):105680, 2024.

- [27] J. J. Heckman and E. J. Vytlačil. Instrumental variables, selection models, and tight bounds on the average treatment effect. In W. Franz, M. Lechner, and F. Pfeiffer, editors, *Econometric Evaluation of Labour Market Policies*, volume 13, pages 1–15. Physica-Verlag HD, Heidelberg, 2001.
- [28] D. Hilbert. Über die stetige abbildung einer linie auf ein flächenstück. *Dritter Band: Analysis-Grundlagen der Mathematik-Physik Verschiedenes: Nebst Einer Lebensgeschichte*, pages 1–2, 1935.
- [29] Y. Hu, Y. Wu, L. Zhang, and X. Wu. A generative adversarial framework for bounding confounded causal effects. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):12104–12112, May 2021.
- [30] Y. Hu, Y. Wu, L. Zhang, and X. Wu. A generative adversarial framework for bounding confounded causal effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12104–12112, 2021.
- [31] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *ArXiv*, abs/1611.01144, 2016.
- [32] N. Kilbertus, M. J. Kusner, and R. Silva. A class of algorithms for general instrumental variable models, Oct. 2020.
- [33] M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017.
- [34] M. Lotfollahi, A. K. Susmelj, C. De Donno, Y. Ji, I. L. Ibarra, F. A. Wolf, N. Yakubova, F. J. Theis, and D. Lopez-Paz. Compositional perturbation autoencoder for single-cell response modeling. *BioRxiv*, 2021.
- [35] G. Luise, A. Rudi, M. Pontil, and C. Ciliberto. Differential Properties of Sinkhorn Approximation for Learning with Wasserstein Distance, May 2018.
- [36] C. J. Maddison, A. Mnih, and Y. W. Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables, Mar. 2017.

- [37] C. F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.
- [38] C. Mao, A. Cha, A. Gupta, H. Wang, J. Yang, and C. Vondrick. Generative interventions for causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2021.
- [39] V. Melnychuk, D. Frauen, and S. Feuerriegel. Partial Counterfactual Identification of Continuous Outcomes with a Curvature Sensitivity Model. *Advances in Neural Information Processing Systems*, 36:32020–32060, Dec. 2023.
- [40] C. H. Miles, P. Kanki, S. Meloni, and E. J. T. Tchetgen. On Partial Identification of the Pure Direct Effect, Sept. 2015.
- [41] K. Padh, J. Zeitler, D. Watson, M. Kusner, R. Silva, and N. Kilbertus. Stochastic causal programming for bounding treatment effects. In *Conference on Causal Learning and Reasoning*, pages 142–176. PMLR, 2023.
- [42] P. Pauli, A. Koch, J. Berberich, P. Kohler, and F. Allgöwer. Training robust neural networks using Lipschitz bounds. *IEEE Control Systems Letters*, 6:121–126, 2021.
- [43] D. Perekrestenko, L. Eberhard, and H. Bölskei. High-Dimensional Distribution Generation Through Deep Neural Networks, Aug. 2022.
- [44] R. R. Ramsahai. Causal bounds and observable constraints for non-deterministic models. *Journal of Machine Learning Research*, 13(29):829–848, 2012.
- [45] A. Richardson, M. G. Hudgens, P. B. Gilbert, and J. P. Fine. Nonparametric Bounds and Sensitivity Analysis of Treatment Effects. *Statistical Science*, 29(4), Nov. 2014.
- [46] M. C. Sachs, G. Jonzon, A. Sjölander, and E. E. Gabriel. A general method for deriving tight symbolic bounds on causal effects, Dec. 2021.
- [47] E. Shchepin. On hölder maps of cubes. *Mathematical Notes*, 87:757–767, 2010.
- [48] J. Tian and J. Pearl. On the testable implications of causal models with hidden variables, Dec. 2002.

- [49] A. Virmaux and K. Scaman. Lipschitz regularity of deep neural networks: Analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.
- [50] K. Xia, K.-Z. Lee, Y. Bengio, and E. Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference, Oct. 2022.
- [51] K. Xia, Y. Pan, and E. Bareinboim. Neural causal models for counterfactual identification and estimation. *arXiv preprint arXiv:2210.00035*, 2022.
- [52] D. Yarotsky. Optimal approximation of continuous functions by very deep relu networks. *ArXiv*, abs/1802.03620, 2018.
- [53] J. Zhang and E. Bareinboim. Bounding causal effects on continuous outcome. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):12207–12215, May 2021.
- [54] J. Zhang and E. Bareinboim. Non-parametric methods for partial identification of causal effects. *Columbia CausalAI Laboratory Technical Report*, 2021.
- [55] J. Zhang, J. Tian, and E. Bareinboim. Partial counterfactual identification from observational and experimental data, Oct. 2021.

A Illustration of Notions in Section 2

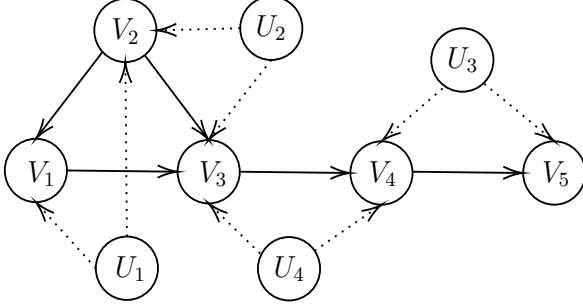


Figure 4: An SCM example.

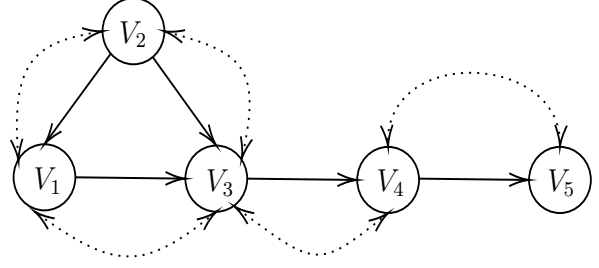


Figure 5: The causal graph of this SCM.

To further explain the notions in Section 2, we consider the following example. Let $\mathcal{M}$ be an SCM with the following structure equations.

$$\begin{aligned}
 V_1 &= f_1(V_2, U_1), \\
 V_2 &= f_2(U_1, U_2), \\
 V_3 &= f_3(V_1, V_2, U_2, U_4), \\
 V_4 &= f_4(V_3, U_3, U_4), \\
 V_5 &= f_5(V_4, U_3),
 \end{aligned} \tag{A.1}$$

where U_1 and U_2 are correlated and U_3, U_4 and (U_1, U_2) are independent. The causal model is shown in Figure 4 and its causal graph is shown in Figure 5. In this example, $\mathbf{U}_{V_1} = \{U_1\}$, $\mathbf{U}_{V_2} = \{U_1, U_2\}$, $\mathbf{U}_{V_3} = \{U_2, U_4\}$, $\mathbf{U}_{V_4} = \{U_3, U_4\}$, $\mathbf{U}_{V_5} = \{U_3\}$. Since U_1, U_2 are correlated, $C_1 = \{V_1, V_2, V_3\}$ is one C^2 component because all nodes in C_1 are connected by bi-directed edges. Note that $\{V_1, V_2, V_3, V_4\}$ is not a C^2 component because V_4 and V_2 is not connected by any bi-directive edge. The rest of C^2 components are $C_2 = \{V_4, V_5\}$, $C_3 = \{V_3, V_4\}$.

Now, we consider the intervention $V_1 = t$. Under this intervention, the structure equations can

be obtained by setting $V_1 = t$ while keeping all other equations unchanged, i.e.,

$$\begin{aligned} V_1(t) &= t, \\ V_2(t) &= f_2(U_1, U_2), \\ V_3 &= f_3(V_1, V_2, U_2, U_4), \\ V_4 &= f_4(V_3, U_3, U_4), \\ V_5(t) &= f_5(V_4, U_3). \end{aligned}$$

The figure of this model under intervention is shown in Figure 6.

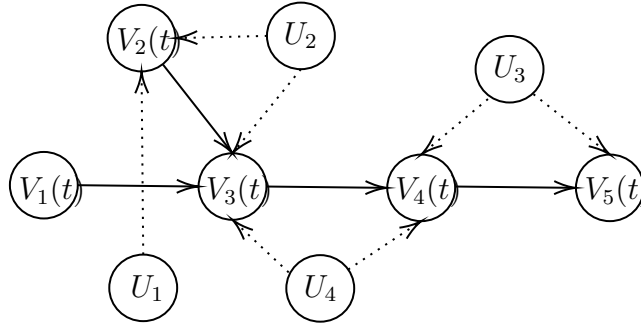


Figure 6: The SCM after intervention $V_1 = t$.

B Proof of approximation Theorems

B.1 Illustration of how to construct canonical representations and neural architectures

In this section, we illustrate how to construct canonical representations and neural architectures given a causal graph via a simple example. We consider the example in the previous section. The causal graph is shown in Figure 5.

Following the construction in Proposition B.1, we use one latent variable for each C^2 component. As we explain in Section A, this causal model has three C^2 components, $\{V_1, V_2, V_3\}$, $\{V_3, V_4\}$, $\{V_4, V_5\}$. In the canonical representation, exactly the latent variables enter their corresponding C^2 component.

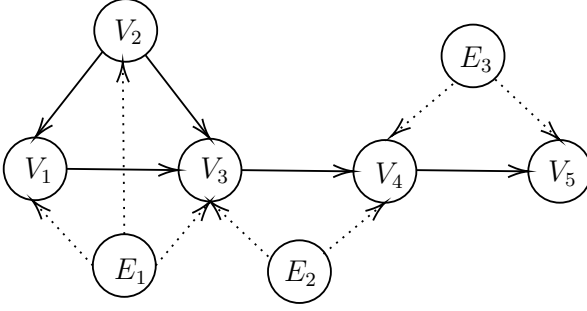


Figure 7: The canonical representation of Figure 5.

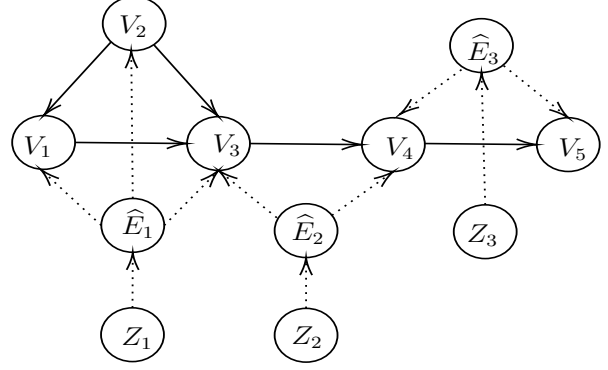


Figure 8: The neural network architecture.

The structure equation of this SCM is as follows.

$$\begin{aligned}
 V_1 &= f_1(V_2, E_1), \\
 V_2 &= f_2(E_1), \\
 V_3 &= f_3(V_1, V_2, E_1, E_2), \\
 V_4 &= f_4(V_3, E_2), \\
 V_5 &= f_5(V_4, E_3),
 \end{aligned} \tag{B.1}$$

If we set $E_1 = (U_1, U_2)$, $E_2 = U_4$, $E_3 = U_3$, (B.1) is equivalent to (A.1). Therefore, we can see from this example that the canonical representation does not lose any information about the SCM.

Now, we show how to construct the NCM architecture from a canonical representation. As we mentioned in Section 3, we approximate the latent distribution by pushing forward uniform and Gumbel variables. The structure equation of the NCM is

$$\begin{aligned}
 V_1 &= f_1^{\theta_1}(V_2, g_1^{\theta_1}(Z_1)), \\
 V_2 &= f_2^{\theta_2}(g_1^{\theta_1}(Z_1)), \\
 V_3 &= f_3^{\theta_2}(V_1, V_2, g_1^{\theta_1}(Z_1), g_2^{\theta_1}(Z_2)), \\
 V_4 &= f_4^{\theta_4}(V_3, g_2^{\theta_1}(Z_2)), \\
 V_5 &= f_5^{\theta_5}(V_4, g_2^{\theta_3}(Z_3)),
 \end{aligned} \tag{B.2}$$

where $f_i^{\theta_i}, g_j^{\theta_j}$ are neural networks, Z_i are joint distribution of independent uniform and Gumbel variables, and $g_j^{\theta_j}$ has the special architecture described in Section 3.1 and Section 3.2. Figure 8 shows the architecture of the NCM. Each in-edge represents an input of a neural net.

B.2 Proof of Theorem 3.3

The causal graph $\mathcal{G}$ does not specify how latent variables influence the observed variables. There could be many ways of recovering the latent variables from a graph $\mathcal{G}$. Figure 9 shows an example where two different causal models have the same causal graph. The next proposition gives a canonical representation of a causal model.

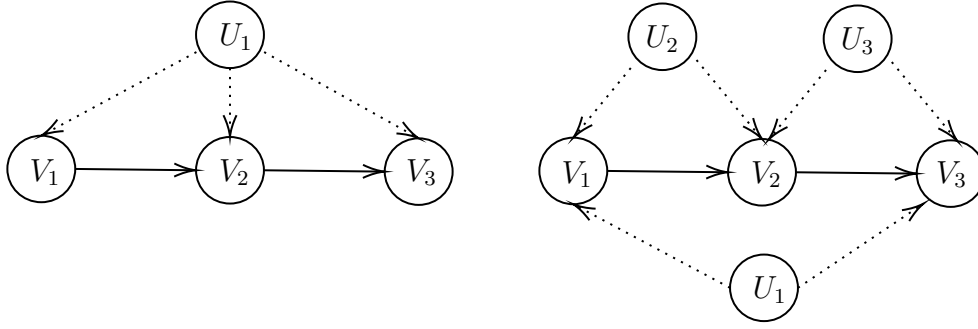


Figure 9: Example: two different SCMs with the same causal graph.

Proposition B.1. *Suppose that ?? 2.3 and ?? 2.4 hold, given any SCM $\mathcal{M}$ with causal graph $\mathcal{G}$ and latent variables $\mathbf{U}$, we can construct a canonical SCM $\widehat{\mathcal{M}}$ of the form (3.1) by merging the latent variables in $\mathbf{U}$ such that $\mathcal{M}$ and $\widehat{\mathcal{M}}$ produce the same intervention results. Besides, the continuous structural functions in $\widehat{\mathcal{M}}$ have the same smoothness as those in $\mathcal{M}$. The categorical functions in $\widehat{\mathcal{M}}$ have the smoothness of the normalized propensity functions $f_i/\|f_i\|_1$: normalization preserves differentiability order and local Lipschitz regularity wherever $\|f_i\|_1 > 0$, and coordinate regrouping preserves any global Lipschitz regularity these normalized functions possess.*

Additional condition. In the categorical structural equations, $\arg \max$ selects the smallest maximizing label when there is a tie. This is a measurable completion on ties; under the stipulated independent standard Gumbel noises and nonnegative propensity weights of positive total, it leaves every fixed-intervention law unchanged.

Remark. The categorical generation formula requires $f_i(x) \in [0, \infty)^{n_i}$ and $\|f_i(x)\|_1 > 0$ on its structural input domain. A zero categorical coordinate has extended logarithm $\log 0 = -\infty$.

Proof of Proposition B.1. Let $\mathbf{G} = \{G_{V_i} : V_i \text{ is categorical}\}$. By ?? 2.3 and ?? 2.4, the remaining latent variables $U_1, \dots, U_{n_U}$ and the Gumbel coordinates are independent. For a categorical variable,

put $s_i(x) = \|f_i(x)\|_1$ and $\pi_i(x) = f_i(x)/s_i(x)$. The categorical generation assumption requires nonnegative coordinates and $s_i(x) > 0$ on its structural input domain. For positive coordinates,

$$\log \pi_{i,j}(x) = \log f_{i,j}(x) - \log s_i(x).$$

The subtracted term is common to every coordinate, so replacing f_i by π_i leaves its Gumbel argmax unchanged. Zero coordinates remain zero and have extended logarithm $-\infty$. Thus this replacement preserves the categorical structural output for every input and the normalized vector belongs to the probability simplex.

Let $D_1, \dots, D_{n_C}$ be the maximal C^2 -components. For a non-Gumbel latent variable define $I(U_k) = \{V_j : U_k \in \mathcal{U}_{V_j}\}$. Independence implies that every nonempty $I(U_k)$ is a bidirected clique, so it is contained in a maximal C^2 -component. Assign U_k to one such component, choosing one when there is more than one choice. Unused latent variables may be assigned arbitrarily. Let $\hat{U}_C$ be the tuple of variables assigned to C ; for an empty tuple use a constant latent variable. Disjoint tuples are independent. Their coordinate projections recover the original joint law of the non-Gumbel latent variables, and they are independent of $\mathcal{G}$.

For each V_i , supply all $\hat{U}_C$ with $V_i \in C$ as canonical inputs. Project these tuples onto the original inputs $\mathcal{U}_{V_i}$ and ignore the other coordinates. Define its continuous function by applying f_i to these projected inputs and its categorical function by applying π_i . Together with the unchanged parent relations and Gumbel noise, these are exactly the normalized structural equations in (3.1). Every original bidirected edge is generated by a shared original latent variable and remains present after that variable is assigned to a component containing both endpoints. Every newly declared shared tuple has its affected vertices inside its clique C , so it introduces no edge outside the original graph. Hence the canonical causal graph is $\mathcal{G}$.

Couple each tuple with its original coordinates and retain the same Gumbel variables. Under any intervention, the treated variables agree by construction. In a topological order, suppose all parents of the next untreated variable agree. Its projected latent inputs agree; its continuous structural function is unchanged, or its categorical argmax agrees by the normalization identity. Its value therefore agrees as well. Induction over this order proves equality of all observed variables under every intervention, and hence equality of their intervention laws.

Coordinate projections preserve differentiability order and are 1-Lipschitz for the 1-norm, so they preserve global Lipschitz constants of the functions to which they are applied. Continuous

functions therefore retain their original regularity, and categorical functions retain the regularity of π_i . To verify the normalization claims, if f_i is differentiable to a given order and $s_i > 0$, the quotient by the positive differentiable function s_i has that same differentiability order. If f_i is locally Lipschitz, fix an input where $s_i > 0$ and a neighborhood on which $s_i \geq a > 0$. For two points x, y in that neighborhood, nonnegativity gives $|s_i(x) - s_i(y)| \leq \|f_i(x) - f_i(y)\|$, and

$$\left\| \frac{f_i(x)}{s_i(x)} - \frac{f_i(y)}{s_i(y)} \right\| \leq \frac{\|f_i(x) - f_i(y)\|}{s_i(x)} + \frac{|s_i(x) - s_i(y)|}{s_i(x)} \leq \frac{2}{a} \|f_i(x) - f_i(y)\|.$$

This proves local Lipschitz regularity of the normalized propensity. If that propensity is globally Lipschitz, composing it with the coordinate projections retains its global bound, completing the stated regularity conclusions. $\square$

Proof of Theorem 3.3. The structure equations of $\mathcal{M}$ have the form

$$V_i = \begin{cases} f_i(\text{Pa}(V_i), \mathbf{U}_{V_i}), & V_i \text{ is continuous,} \\ \arg \max_{k \in [n_i]} \{g_k^{V_i} + \log(f_i(\text{Pa}(V_i), \mathbf{U}_{V_i}))_k\}, & V_i \text{ is categorical, } \|f_i\|_1 = 1, \end{cases}$$

where $g_k^{V_i}$ are i.i.d. standard Gumbel variables and $\widehat{\mathcal{M}} \in \mathcal{M}(\mathcal{G}, \widehat{\mathcal{F}}, \widehat{\mathbf{U}})$ has structure equations

$$\widehat{V}_i = \begin{cases} \widehat{f}_i(\text{Pa}(\widehat{V}_i), \widehat{\mathbf{U}}_{V_i}), & V_i \text{ is continuous,} \\ \arg \max_{k \in [n_i]} \{\widehat{g}_k^{V_i} + \log(\widehat{f}_i(\text{Pa}(\widehat{V}_i), \widehat{\mathbf{U}}_{V_i}))_k\}, & V_i \text{ is categorical, } \|\widehat{f}_i\|_1 = 1, \end{cases}$$

where $\widehat{g}_k^{V_i}$ are i.i.d. standard Gumbel variables. Let the treatment variables set be $\mathbf{T} = \{T_1, \dots, T_{n_T}\}$. We may give all vertices and $\mathbf{U}_0 = \{U_1, \dots, U_{n_U}\}$, an topology ordering (rearrange the subscript if necessary)

$$U_1, \dots, U_{n_U}, T_1, \dots, T_{n_T}, V_1(\mathbf{t}), \dots, V_{n_V - n_T}(\mathbf{t})$$

such that for each directed edge $(V_i(\mathbf{t}), V_j(\mathbf{t}))$, vertex $V_i(\mathbf{t})$ lies before vertex $V_j(\mathbf{t})$ in the ordering, ensuring that all edges start from vertices that appear early in the order and end at vertices that appear later in the order. We put $\mathbf{U}_0$ and $\mathbf{T}$ at the beginning because they are the root nodes of the intervened model. We denote $\mu_{\mathbf{U}_0, \mathbf{T}, 1:k}^T$ (resp. $\widehat{\mu}_{\mathbf{U}_0, \mathbf{T}, 1:k}^T$) to be the distribution of $\mathbf{U}_0, \mathbf{T}, V_1(\mathbf{t}), \dots, V_k(\mathbf{t})$ (resp. $\widehat{\mathbf{U}}_0, \mathbf{T}, \widehat{V}_1(\mathbf{t}), \dots, \widehat{V}_k(\mathbf{t})$), $\mu_{V_k|\mathbf{U}_0, \mathbf{T}, 1:k-1}^T$ the distribution of V_k given $\mathbf{U}_0, \mathbf{T}, V_1(\mathbf{t}), \dots, V_{k-1}(\mathbf{t})$.

Remark (notation). $\widehat{\mu}$ denotes the comparison probability measure on the same space as μ . In Lemma E.1: $\widehat{\mu}$ is supported on the compact set $K \subset \mathbb{R}^{d_1}$. In Lemma E.3: $\widehat{\mu} = \widehat{\mu}_1 \otimes \dots \otimes \widehat{\mu}_n$ is the product of the comparison factor measures.

Let $S = \sum_{i=1}^{n_V} \|f_i - \widehat{f}_i\|_\infty$. Put

$$D = \max(\{0\} \cup \{n_i - 1 : V_i \text{ is categorical}\}),$$

$$A = \max\{1+L, 1+LD/2\}, \quad B = \max(\{1\} \cup \{d_i : V_i \text{ is continuous}\} \cup \{n_i(n_i - 1)/2 : V_i \text{ is categorical}\}).$$

Write $x = (\mathbf{u}, \mathbf{t}, v_1, \dots, v_{k-1})$ for a history, and let E_k denote the Wasserstein distance between the two history laws through V_k . Next, we prove that

$$W(\mu_{\mathbf{U}_0, \mathbf{T}, 1:k}^T, \widehat{\mu}_{\widehat{\mathbf{U}}_0, \mathbf{T}, 1:k}^T) \leq AW(\mu_{\mathbf{U}_0, \mathbf{T}, 1:k-1}^T, \widehat{\mu}_{\widehat{\mathbf{U}}_0, \mathbf{T}, 1:k-1}^T) + BS. \quad (\text{B.3})$$

By definition of Wasserstein-1 distance,

$$\begin{aligned} W(\mu_{\mathbf{U}_0, \mathbf{T}, 1:k}^T, \widehat{\mu}_{\widehat{\mathbf{U}}_0, \mathbf{T}, 1:k}^T) &= \sup_{g \in \text{Lip}(1)} \int g(\mathbf{u}, \mathbf{t}, v_1, \dots, v_k) d(\mu_{\mathbf{U}_0, \mathbf{T}, 1:k}^T - \widehat{\mu}_{\widehat{\mathbf{U}}_0, \mathbf{T}, 1:k}^T) \\ &= \sup_{g \in \text{Lip}(1)} \int g(\mathbf{u}, \mathbf{t}, v_1, \dots, v_k) d\mu_{V_k|\mathbf{U}_0, \mathbf{T}, 1:k-1}^T d\mu_{\mathbf{U}_0, \mathbf{T}, 1:k-1}^T \\ &\quad - \int g(\mathbf{u}, \mathbf{t}, v_1, \dots, v_k) d\widehat{\mu}_{\widehat{V}_k|\widehat{\mathbf{U}}_0, \mathbf{T}, 1:k-1}^T d\widehat{\mu}_{\widehat{\mathbf{U}}_0, \mathbf{T}, 1:k-1}^T \\ &= \sup_{g \in \text{Lip}(1)} \underbrace{\int g(\mathbf{u}, \mathbf{t}, v_1, \dots, v_k) d(\mu_{V_k|\mathbf{U}_0, \mathbf{T}, 1:k-1}^T - \widehat{\mu}_{\widehat{V}_k|\widehat{\mathbf{U}}_0, \mathbf{T}, 1:k-1}^T) d\widehat{\mu}_{\widehat{\mathbf{U}}_0, \mathbf{T}, 1:k-1}^T}_{(1)} \\ &\quad + \underbrace{\int g(\mathbf{u}, \mathbf{t}, v_1, \dots, v_k) d\mu_{V_k|\mathbf{U}_0, \mathbf{T}, 1:k-1}^T d(\mu_{\mathbf{U}_0, \mathbf{T}, 1:k-1}^T - \widehat{\mu}_{\widehat{\mathbf{U}}_0, \mathbf{T}, 1:k-1}^T)}_{(2)}. \end{aligned}$$

For (1), if V_k is continuous, the two conditional laws at history x are point masses at $f_k(x)$ and $\widehat{f}_k(x)$. Since g is 1-Lipschitz in the paper's 1-norm,

$$\begin{aligned} \left| \int g(x, v_k) d(\mu_{V_k|x}^T - \widehat{\mu}_{\widehat{V}_k|x}^T) \right| &= |g(x, f_k(x)) - g(x, \widehat{f}_k(x))| \\ &\leq \|f_k(x) - \widehat{f}_k(x)\| \\ &\leq d_k \|f_k - \widehat{f}_k\|_\infty \leq BS. \end{aligned} \quad (\text{B.4})$$

If V_k is categorical, write $p(x) = f_k(x)$ and $\widehat{p}(x) = \widehat{f}_k(x)$, with n_k coordinates. These vectors belong to Δ_{n_k} . Put $c(x) = (\max_j g(x, j) + \min_j g(x, j))/2$. The range of $j \mapsto g(x, j)$ is at most

$n_k - 1$, and $\sum_j (p_j(x) - \hat{p}_j(x)) = 0$. Therefore

$$\begin{aligned} \left| \sum_{j=1}^{n_k} (p_j(x) - \hat{p}_j(x)) g(x, j) \right| &= \left| \sum_{j=1}^{n_k} (p_j(x) - \hat{p}_j(x)) (g(x, j) - c(x)) \right| \\ &\leq \frac{n_k - 1}{2} \sum_{j=1}^{n_k} |p_j(x) - \hat{p}_j(x)| \\ &\leq \frac{n_k(n_k - 1)}{2} \|f_k - \hat{f}_k\|_\infty \leq BS. \end{aligned} \quad (\text{B.5})$$

Remark (notation). f_k denotes a locally indexed deterministic map whose domain and meaning are those of the surrounding structural-model or partition construction. In Theorem 3.3, Theorem 4.2: f_k is the structural function at observed node V_k , with inputs its parents and affecting non-Gumbel latent variables; its output is the continuous value or the normalized categorical propensity vector, as appropriate.

For (2), if V_i is continuous,

$$\begin{aligned} &\int g(\mathbf{u}, \mathbf{t}, v_1, \dots, v_k) d\mu_{V_k|U_0, \mathbf{T}, 1:k-1}^T d\left(\mu_{U_0, \mathbf{T}, 1:k-1}^T - \hat{\mu}_{U_0, \mathbf{T}, 1:k-1}^T\right) \\ &= \int g(\mathbf{u}, \mathbf{t}, v_1, \dots, v_{k-1}, f_k(\text{pa}(v_k), \mathbf{u}_{V_k})) d\left(\mu_{U_0, \mathbf{T}, 1:k-1}^T - \hat{\mu}_{U_0, \mathbf{T}, 1:k-1}^T\right). \end{aligned} \quad (\text{B.6})$$

Since g, f_k are Lipschitz continuous functions, $g(\mathbf{u}, \mathbf{t}, v_1, \dots, v_{k-1}, f_k(\text{pa}(v_k), \mathbf{u}_{V_k}))$ is $(L + 1)$ -Lipschitz continuous with respect to $(\mathbf{u}, \mathbf{t}, v_1, \dots, v_{k-1})$. We have

$$\begin{aligned} &\int g(\mathbf{u}, \mathbf{t}, v_1, \dots, v_{k-1}, f_k(\text{pa}(v_k), \mathbf{u}_{V_k})) d\left(\mu_{U_0, \mathbf{T}, 1:k-1}^T - \hat{\mu}_{U_0, \mathbf{T}, 1:k-1}^T\right) \\ &\leq (L + 1)W \left(\mu_{U_0, \mathbf{T}, 1:k-1}^T, \hat{\mu}_{U_0, \mathbf{T}, 1:k-1}^T\right). \end{aligned} \quad (\text{B.7})$$

If V_k is categorical, the true conditional expectation of the test function is $H(x) = \sum_{j=1}^{n_k} p_j(x) g(x, j)$.

For two histories x, y , subtracting and adding $\sum_j p_j(x) g(y, j)$ gives

$$H(x) - H(y) = \sum_j p_j(x) (g(x, j) - g(y, j)) + \sum_j (p_j(x) - p_j(y)) g(y, j).$$

The first sum has absolute value at most $\|x - y\|$. In the second sum subtract the midpoint of the range of $j \mapsto g(y, j)$; the probabilities sum to one and that range has length at most $n_k - 1$. Because the propensity vector is L -Lipschitz in the 1-norm,

$$|H(x) - H(y)| \leq \|x - y\| + \frac{n_k - 1}{2} \|p(x) - p(y)\| \leq \left(1 + \frac{(n_k - 1)L}{2}\right) \|x - y\| \leq A \|x - y\|.$$

It follows from the Wasserstein dual bound that

$$\left| \int H d \left(\mu_{\mathbf{U}_0, \mathbf{T}, 1:k-1}^{\mathbf{T}} - \hat{\mu}_{\hat{\mathbf{U}}_0, \mathbf{T}, 1:k-1}^{\mathbf{T}} \right) \right| \leqslant A E_{k-1}. \quad (\text{B.8})$$

Combine (B.4)-(B.8), we prove Equation (B.3). The recurrence gives, for $m = n_V - n_T$,

$$E_m \leqslant A^m E_0 + BS \sum_{r=0}^{m-1} A^r.$$

Indeed, at $m = 0$ the sum is empty and this is equality. If it holds at m , then

$$E_{m+1} \leqslant A E_m + BS \leqslant A^{m+1} E_0 + BS \left(1 + \sum_{r=1}^m A^r \right) = A^{m+1} E_0 + BS \sum_{r=0}^m A^r.$$

This induction does not divide by L and includes $L = 0$. Define

$$C_{\mathcal{G}}(L, K; \{d_i, n_i\}) = \max \left\{ A^{n_V}, B \sum_{r=0}^{n_V-1} A^r \right\}.$$

The identical intervention values contribute zero distance. The independent Gumbel coordinates have the same law in the two models: coupling them identically gives an upper bound by the Wasserstein distance of the non-Gumbel latent laws, while projection onto those latents gives the reverse bound. Thus

$$E_0 = W \left(P^{\mathcal{M}}(\mathbf{U}), P^{\hat{\mathcal{M}}}(\hat{\mathbf{U}}) \right).$$

Projection from the joint history to the observed variables is 1-Lipschitz, and $m \leqslant n_V$. Consequently

$$\begin{aligned} W \left(P^{\mathcal{M}}(\mathbf{V}(t)), P^{\hat{\mathcal{M}}}(\hat{\mathbf{V}}(t)) \right) &\leqslant E_m \\ &\leqslant C_{\mathcal{G}}(L, K; \{d_i, n_i\}) \left(S + W \left(P^{\mathcal{M}}(\mathbf{U}), P^{\hat{\mathcal{M}}}(\hat{\mathbf{U}}) \right) \right). \end{aligned} \quad (\text{B.9})$$

□

B.3 Proof of results in Section 3.1: Approximating with Wide NN

Proof of Theorem 3.6 Part I. Put $p_i = \mathbb{P}(C_i)$ and $\mu_i = \mathbb{P}|_{C_i}/p_i$. Each component is compact because it is homeomorphic to a compact cube. Distinct components have positive distance because

they are disjoint compact sets and there are finitely many of them. A sufficiently small neighborhood of a support point in C_i meets no other component and has positive measure; hence $p_i > 0$. If $d_i^C = 0$, write $C_i = \{c_i\}$, use the unique map $\widehat{g}_1^i : [0, 1] \rightarrow [0, 1]^0$, and set $\widehat{g}_2^i = c_i$. This produces μ_i exactly.

For $q = d_i^C \geq 1$, choose $g_2^i \in \mathcal{H}([0, 1]^q, C_i) \cap \mathcal{F}_L([0, 1]^q, C_i)$ from Assumption 3.4, and put $\nu_i = ((g_2^i)^{-1})_{\#} \mu_i$. We instantiate Theorem VIII.1 of [43] with its positive integer depth parameter $s = 1$. For a refinement integer $n \geq 2$, set

$$\delta = \frac{1}{\lceil \sqrt{q}(q+1)n(n-1) \rceil}.$$

Since $\sqrt{q}(q+1) \geq 2$, this choice satisfies

$$0 < \delta \leq \frac{1}{\sqrt{q}(q+1)n(n-1)} < \frac{1}{n(n-1)}.$$

The quantization mesh in the cited theorem is δ/n . Its conclusion supplies a one-input, q -output ReLU network Φ with at most $C_q(n^q + n^{q-1})$ nonzero weights and biases, depth $4q - 2$ in the source's convention, and Euclidean Wasserstein error at most

$$\frac{\sqrt{q}}{2n} + \frac{2\sqrt{q}}{n} + \frac{q(q+1)}{2}(n-1)\delta \leq \frac{\sqrt{q}}{2n} + \frac{2\sqrt{q}}{n} + \frac{\sqrt{q}}{2n} = \frac{3\sqrt{q}}{n}.$$

Here C_q depends only on q . The pointwise inequality between the 1-norm and the Euclidean norm multiplies this Wasserstein bound by at most $\sqrt{q}$, giving error at most $3q/n$ in the paper's norm.

A hidden unit with zero incoming weights and zero bias has identically zero output and can be deleted. After deleting these units, every remaining hidden unit accounts for a nonzero weight or bias, so every hidden layer has width at most $C_q(n^q + n^{q-1}) \leq 2C_q n^q$. Clip the output coordinates into $[0, 1]$ using $t \mapsto \max\{t, 0\} - \max\{t - 1, 0\}$. This requires width $2q$ and a fixed number of additional layers. Clipping fixes the target cube and is 1-Lipschitz in the 1-norm, so its application to the first marginal of any coupling cannot increase the error. There is consequently a constant $A_q \geq 1$, depending only on q , such that the clipped network has width at most $A_q n^q$ and hidden depth $\Theta(q)$.

For every sufficiently large width budget, specifically $W_1 \geq A_q 2^q$, choose

$$n = \lfloor (W_1/A_q)^{1/q} \rfloor.$$

Then $n \geq 2$, $A_q n^q \leq W_1$, and $n \geq \frac{1}{2}(W_1/A_q)^{1/q}$. With this choice the clipped network $\widehat{g}_1^i$ satisfies

$$W(P(\widehat{g}_1^i(U_i)), \nu_i) \leq \frac{3q}{n} \leq 6q A_q^{1/q} W_1^{-1/q}, \quad U_i \sim \text{Unif}([0, 1]).$$

Thus the refinement integer increases with width, while the depth parameter stays equal to 1; the mesh δ/n is $\Theta_q(W_1^{-3/q})$. Padding with zero units realizes width W_1 without altering the function.

For each scalar coordinate of g_2^i , apply Theorem 2(b) of [52] with positive input dimension q . It supplies width $2q + 10$ and error bounded by a constant times the coordinate modulus of continuity at a constant times the number of weights raised to $-2/q$. At this fixed width the number of weights is proportional to depth. The Lipschitz modulus therefore gives coordinate error $O(L_2^{-2/q})$. Parallelizing the d coordinate networks gives width $\Theta(dq)$ and a map $\widehat{g}_2^i$ with $\|g_2^i - \widehat{g}_2^i\|_\infty = O(L_2^{-2/q})$. Reserve a fixed number of layers for clipping into a box $[-B, B]^d$ containing all components; the remaining depth is proportional to L_2 for sufficiently large L_2 , and clipping cannot increase any coordinate error. Lemma E.1 gives

$$W(\mu_i, P(\widehat{g}_2^i \circ \widehat{g}_1^i(U_i))) \leq L_i O(W_1^{-1/q}) + d O(L_2^{-2/q}) = O(W_1^{-1/q} + L_2^{-2/q}),$$

where L_i is the fixed chart Lipschitz constant.

Take independent uniform inputs U_i . Pad the first-block hidden depths to a common value $\Theta(D_C)$ and the second-block widths to $\Theta(dD_C)$. Identity layers on nonnegative hidden outputs preserve the maps. Let $\widehat{\mu}_i = P(\widehat{g}_2^i \circ \widehat{g}_1^i(U_i))$. For budgets at least 1, each positive $q \leq D_C$ satisfies $W_1^{-1/q} \leq W_1^{-1/D_C}$ and $L_2^{-2/q} \leq L_2^{-2/D_C}$. Singleton errors are zero. Lemma E.2 hence yields

$$W\left(\mathbb{P}, \sum_i p_i \widehat{\mu}_i\right) \leq \sum_i p_i W(\mu_i, \widehat{\mu}_i) \leq O\left(\mathbb{I}_{\{\max_i d_i^C > 0\}}(W_1^{-1/D_C} + L_2^{-2/D_C})\right).$$

Let X^τ be the Gumbel-softmax weights with probabilities p_i , independent of the block outputs, and let X^0 be their one-hot limit. The output with weights X^0 has law $\sum_i p_i \widehat{\mu}_i$. Lemma E.4 gives $W(P(X^\tau), P(X^0)) = O(\tau - \tau \log \tau)$ for $0 < \tau < 1$. Couple the weights and use identical independent block outputs v_i in the two constructions. Since $\|v_i\|_\infty \leq B$,

$$\left\| \sum_i (X_i^\tau - X_i^0) v_i \right\| \leq \sum_i |X_i^\tau - X_i^0| \|v_i\| \leq dB \|X^\tau - X^0\|.$$

Taking expectations and the infimum over weight couplings bounds the additional output Wasserstein error by $O(\tau - \tau \log \tau)$. The triangle inequality proves the stated estimate. If every component is a singleton, both deterministic blocks are exact and only the mixing error remains. $\square$

Remark (notation). μ_i denotes the component measure or probability-vector coordinate specified by its local decomposition. In Theorem 3.6: $i \in [N_C]$ and $\mu_i = \mathbb{P}_{C_i}/\mathbb{P}(C_i)$, where $\mathbb{P}_{C_i}$ is the restriction of $\mathbb{P}$ to C_i , whenever $\mathbb{P}(C_i) > 0$. In Lemma E.2: $i \in [n]$ and μ_i is an arbitrary measure on $\mathbb{R}^d$ entering the mixture $\sum_{i=1}^n p_i \mu_i$. In Lemma E.3: $i \in [n]$ and μ_i is the i th factor of $\mu = \mu_1 \otimes \cdots \otimes \mu_n$, paired with $\hat{\mu}_i$ on its factor space. In the result split from Theorem 3.6: $i \in [N_C]$ and $\mu_i = \mathbb{P}_{C_i}/\mathbb{P}(C_i)$, where $\mathbb{P}_{C_i}$ is the restriction of $\mathbb{P}$ to C_i , whenever $\mathbb{P}(C_i) > 0$. In Lemma E.5: $i \in [n]$ and $\mu_i \in [0, 1]$ is the mass at x_i in the finite probability measure with weight vector $\mu \in \Delta_n$; a transportation plan satisfies $\sum_j T_{ij} = \mu_i$. In a result cited in the proof of Theorem 3.6: $i \in [N_C]$ and $\mu_i = \mathbb{P}_{C_i}/\mathbb{P}(C_i)$, where $\mathbb{P}_{C_i}$ is the restriction of $\mathbb{P}$ to C_i , whenever $\mathbb{P}(C_i) > 0$. For $i \in [N_C]$, g_2^i is a chosen map in $\mathcal{H}([0, 1]^{d_i^C}, C_i) \cap \mathcal{F}_L([0, 1]^{d_i^C}, C_i)$, whose inverse pulls the component law back to the unit cube. In the result split from Theorem 3.6: g_2^i is chosen so that $(g_2^i)^{-1}_{\#} \mu_i(B) \geq C_g \lambda(B)/\mathbb{P}(C_i)$ for every measurable $B \subset [0, 1]^{d_i^C}$, where $\mu_i = \mathbb{P}_{C_i}/\mathbb{P}(C_i)$ and λ is cube Lebesgue measure. In Theorem 3.6, the result split from Theorem 3.6: For $d_i^C = 0$, g_2^i is the unique map from $[0, 1]^0$ to the singleton C_i . For $i \in [N_C]$ and $j \in [d]$, $\hat{g}_2^{i,j} : [0, 1]^{d_i^C} \rightarrow \mathbb{R}$ denotes a coordinate network chosen to approximate $(g_2^i)_j$ uniformly on $[0, 1]^{d_i^C}$. In Theorem 3.6: When $d_i^C = 0$, $\hat{g}_2^{i,j}$ is the j th coordinate of the constant affine map to the singleton C_i , so its approximation error is zero. h denotes the auxiliary function or positive grid spacing defined in its surrounding passage. In Theorem 3.6: $h(w, X) = w^\top X \in \mathbb{R}^d$, where w has N_C coordinates, X collects the N_C component outputs, and the stated region is $\|w\|_1 = 1$, $\|X\|_\infty \leq K$. In Lemma C.3: $h(t) = (\hat{f}(t) - f(t))\hat{f}(t)$ is the real-valued auxiliary function in the proof, with the stated Lipschitz constant $L_h = 4LK$. In Lemma E.5: $h(T) = 1 - \sum_{i=1}^n \sum_{j=1}^m T_{ij} \log T_{ij}$ is the entropy expression for $T \in \Pi(\mu, \nu)$, with $0 \log 0 = 0$. In Lemma E.6: h is a scalar test function in $\mathcal{H} = \{h : [-K, K]^d \rightarrow \mathbb{R} : h(0) = 0, |h(x) - h(y)| \leq \|x - y\|\}$, so its values lie in $[-dK, dK]$; the concentration argument first fixes such an h and uses $z_f = \sqrt{(h \circ f + dK)/(2dK)}$. In Theorem 4.2, Corollary 4.4: In the structural interpolation construction, $h = 2K/R > 0$ is the side length of a grid cell obtained by dividing each coordinate of $[-K, K]^m$ into R equal intervals; $I_h f$ is the resulting piecewise affine interpolant. In Proposition 4.3: In the proof, $s = (a, t)$ and $h(s) = (\hat{f}_y(s) - f_y(s))\hat{f}_y(s)$ on the bounded covariate domain $\mathcal{Q}$. The structural regression versions

satisfy $|f_y|, |\widehat{f}_y| \leq B$ there, and $|h(s) - h(s')| \leq 4BL_y\|s - s'\|$. For $i \in [N_C]$, $\widehat{g}_1^i$ denotes the chosen first generator block for component C_i , with scalar input and d_i^C output coordinates. In the result split from Theorem 3.6: For $d_i^C > 0$, $\widehat{g}_1^i$ is the coordinatewise deep ReLU approximation of the curve $\gamma_i : [0, 1] \rightarrow [0, 1]^{d_i^C}$ satisfying $(\gamma_i)_\# \lambda = (g_2^i)^{-1}_\# \mu_i$; its coordinate networks have width 12 and depth L_1 , its combined width is $\Theta(d_i^C)$, and $\|\widehat{g}_1^i - \gamma_i\|_\infty \leq O(L_1^{-2/d_i^C})$. In Theorem 3.6, the result split from Theorem 3.6: For $d_i^C = 0$, $\widehat{g}_1^i$ is the unique map to $[0, 1]^0$. In Theorem 3.6: For $d_i^C > 0$, $\widehat{g}_1^i$ is a quantized ReLU network of width W_1 and depth $\Theta(d_i^C)$, chosen so that $W((g_2^i)^{-1}_\# \mu_i, P(\widehat{g}_1^i(U_i))) \leq O(W_1^{-1/d_i^C})$. In a result cited in the proof of Theorem 3.6: For a component of positive dimension $q = d_i^C$, $\widehat{g}_1^i$ is the quantized ReLU component generator of width W_1 and depth $\Theta(q)$ satisfying $W((g_2^i)^{-1}_\# \mu_i, P(\widehat{g}_1^i(U_i))) \leq O(W_1^{-1/q})$. The source construction uses integer depth parameter $s = 1$, quantization parameter $\delta = 1/\lceil \sqrt{q}(q+1)n(n-1) \rceil$, and mesh δ/n , where $n = \lfloor (W_1/A_q)^{1/q} \rfloor$ for $W_1 \geq A_q 2^q$ and $A_q \geq 1$ depends only on q .

Proof of Corollary 3.7. Put $p_i = \mathbb{P}(C_i)$ and $\mu_i = \mathbb{P}|_{C_i}/p_i$. Each p_i is positive: the finitely many compact support components are separated, and a neighborhood meeting only C_i has positive measure by the definition of support. If $d_i^C = 0$, write $C_i = \{c_i\}$, use the unique map $\widehat{g}_1^i : [0, 1] \rightarrow [0, 1]^0$, and set $\widehat{g}_2^i = c_i$; the component error is zero. For $q = d_i^C \geq 1$, choose the chart $g_2^i : [0, 1]^q \rightarrow C_i$ supplied by Assumption 3.5. Its pullback measure $((g_2^i)^{-1})_\# \mu_i$ satisfies the cube version of that assumption. Proposition B.2 supplies a curve $\gamma_i : [0, 1] \rightarrow [0, 1]^q$ with this law and coordinate moduli bounded by a constant times $r^{1/q}$. Apply Theorem 2(b) of [52] to each coordinate of this one-input curve. Its modulus estimate gives uniform coordinate error $O(L_1^{-2/q})$. Parallelizing the q coordinate networks gives width $\Theta(q)$; clipping their outputs into $[0, 1]^q$ adds a fixed number of layers and cannot increase the errors. Coupling the two outputs at the same uniform input gives

$$W(P(\widehat{g}_1^i(U_i)), ((g_2^i)^{-1})_\# \mu_i) \leq q \|\widehat{g}_1^i - \gamma_i\|_\infty = O(L_1^{-2/q}).$$

For the second block, repeat the chart-coordinate approximation construction in the proof of Theorem 3.6. In that construction, Theorem 2(b) of [52] applies with positive input dimension q to each Lipschitz coordinate of g_2^i . It supplies coordinate width $2q + 10$ and error $O(L_2^{-2/q})$. Parallelizing the d coordinates gives width $\Theta(dq)$ and uniform infinity error $O(L_2^{-2/q})$. Lemma E.1 therefore gives

$$W(\mu_i, P(\widehat{g}_2^i \circ \widehat{g}_1^i(U_i))) \leq L_i O(L_1^{-2/q}) + d O(L_2^{-2/q}) = O(L_1^{-2/q} + L_2^{-2/q}),$$

where L_i is the fixed chart Lipschitz constant. Take independent uniform inputs U_i , and clip second-block outputs into a fixed box $[-B, B]^d$ containing all components; this cannot increase their uniform errors. Padding the parallel blocks gives the stated budgets. Let $\hat{\mu}_i = P(\hat{g}_2^i \circ \hat{g}_1^i(U_i))$. Lemma E.2 gives

$$W\left(\mathbb{P}, \sum_i p_i \hat{\mu}_i\right) \leq \sum_i p_i W(\mu_i, \hat{\mu}_i) \leq O\left(\mathbb{I}_{\{\max_i d_i^C > 0\}}(L_1^{-2/D_C} + L_2^{-2/D_C})\right).$$

Use Gumbel-softmax weights X^τ with probabilities p_i , independent of all block outputs. Their zero-temperature weights X^0 select the mixture $\sum_i p_i \hat{\mu}_i$. By Lemma E.4, $W(P(X^\tau), P(X^0)) = O(\tau - \tau \log \tau)$ for $0 < \tau < 1$. Couple these weights and use identical independent block outputs in the two constructions. For any such outputs $v_i \in [-B, B]^d$,

$$\left\| \sum_i (X_i^\tau - X_i^0) v_i \right\| \leq dB \|X^\tau - X^0\|.$$

Taking expectations and then the infimum over weight couplings bounds the additional output Wasserstein error by $O(\tau - \tau \log \tau)$. The triangle inequality proves the displayed result, including the case where every component is a singleton.

The width factors here can be chosen universally. For $q \geq 1$, each first-block scalar network has width 12, so parallelizing its q coordinates uses width $12q$. Each second-block scalar network has width $2q + 10 \leq 12q$, so its d coordinates use width at most $12dq$. The first-block clipping uses at most $2q$ units, the second-block clipping at most $2d$ units, and carrying a signed intermediate coordinate uses two units. Thus reserving four layers for these operations gives widths bounded by $64q$ and $64dq$, respectively. Each remaining budget is at least half the corresponding total budget once that total budget is at least eight, preserving each displayed error order. A singleton first block is the unique map to the zero-dimensional cube; its constant second block can be carried by at most $2d$ units. Since $D_C \geq 1$, all component blocks therefore fit widths $64D_C$ and $64dD_C$. The architecture is separable, so these are bounds for each component block and are independent of the number of components. Zero units pad widths, and identity layers on nonnegative hidden outputs pad depths; affine output maps recover signed coordinates. When $d = 0$, the target and every component are the unique point of $\mathbb{R}^0$, and the zero-dimensional blocks are exact. This proves the absolute-factor interpretation without changing any generated law. $\square$

B.4 Proof of results in Section 3.1: Approximating with Deep NN

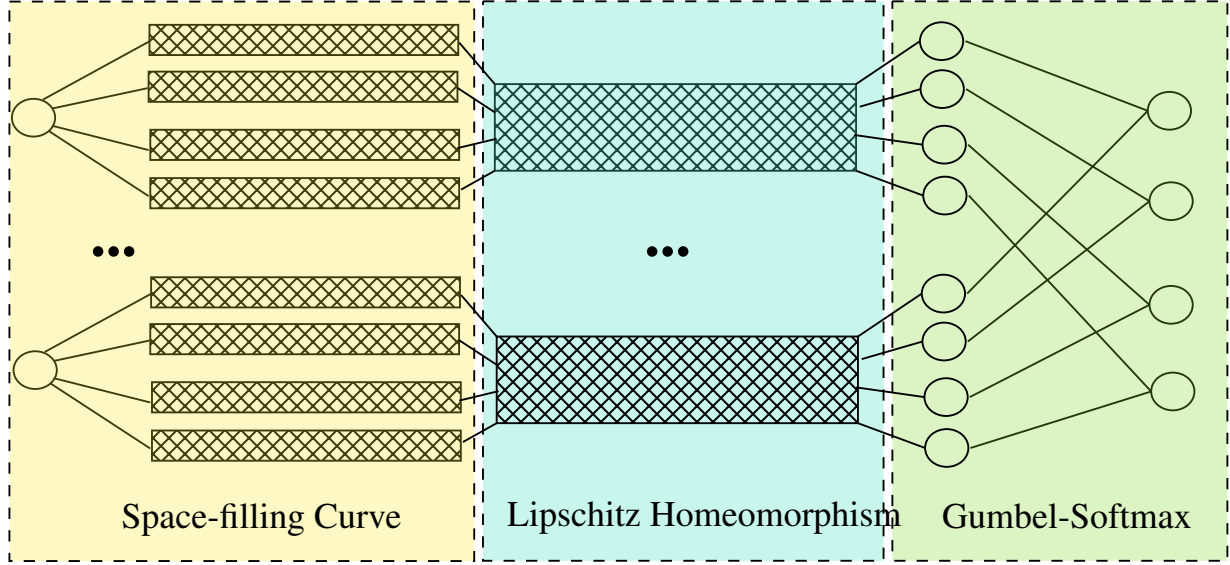


Figure 10: Architecture of the deep neural network for 4-dimensional output. The first (yellow) part approximates the distribution on different connected components using deep ReLu Networks. The remaining two parts are similar to the wide neural network in Figure 1.

We first start with an important proposition.

Proposition B.2. *Let λ be the Lebesgue measure on $[0, 1]$. Given any probability measure $\mathbb{P}$ that satisfies ?? 3.5, there exists a continuous curve $\gamma : [0, 1] \rightarrow \text{supp}(\mathbb{P})$ such that $\gamma_{\#}\lambda = \mathbb{P}$. Furthermore, if $d \geq 1$, γ is $1/d$ -Hölder continuous.*

Remark. For every n , the cells of $\mathcal{C}_n$ are closed coordinate boxes covering $[0, 1]^d$; a cell may have zero side lengths. For distinct cells $C, C' \in \mathcal{C}_n$, the partition condition is $\text{int}(C) \cap \text{int}(C') = \emptyset$, with interiors taken in $\mathbb{R}^d$. In the proof of Proposition B.2, $\mathcal{C}_n$ denotes the standard dyadic collection $\left\{ \prod_{r=1}^d [a_r 2^{-n}, (a_r + 1) 2^{-n}] : a_r \in \{0, \dots, 2^n - 1\} \right\}$. The reference measure λ_d denotes Lebesgue measure on $[0, 1]^d$. In the proof of Proposition B.2, $\mathcal{I}_n$ denotes $\{I_{O_n(i)} : 1 \leq i \leq 2^{nd}\}$ with $I_{O_n(i)} = [(i - 1) 2^{-nd}, i 2^{-nd}]$. The mass assigned to $I_{O_n(i)}$ is $\lambda_d(C_{O_n(i)}) = 2^{-nd}$.

Remark (notation). γ denotes a continuous measure-representing curve parametrized by $[0, 1]$, with parameter measure λ equal to Lebesgue measure on $[0, 1]$. In Proposition B.2: In the statement, $\gamma : [0, 1] \rightarrow \text{supp}(\mathbb{P})$ satisfies $\gamma_{\#}\lambda = \mathbb{P}$ and is $1/d$ -Hölder continuous when $d \geq 1$. In the proof's

change of coordinates, the locally denoted γ instead satisfies $\gamma_{\#}\lambda = f_{\#}^{-1}\mathbb{P}$; the curve representing the original measure is $f \circ \gamma$. For $d = 0$, the representing curve is constant.

Before we prove Proposition B.2, we need to introduce some important notions. Let $\{\mathcal{C}_n\}$ be a sequence of partitions of the unit cube $[0, 1]^d$. We say $\{\mathcal{C}_n\}$ has property (*) if the following properties are satisfied.

1. $\mathcal{C}_0 = \{C_{-1}\}, C_{-1} = [0, 1]^d$
2. $\mathcal{C}_{n+1} = \{C_{i_1, i_2, \dots, i_{n+1}}\}_{i_j=0, \dots, 2^d-1}$ is a refinement of $\mathcal{C}_n = \{C_{i_1, i_2, \dots, i_n}\}_{i_j=0, \dots, 2^d-1}$, i.e., $C_{i_1, i_2, \dots, i_n} = \bigcup_{i_{n+1}=0}^{2^d-1} C_{i_1, i_2, \dots, i_{n+1}}$. Besides, $\mathcal{C}_n$ is obtained by cutting the unit cube $[0, 1]^d$ by planes that are paralleled to the coordinate planes.

Remark (notation). For $n \geq 0$, $\mathcal{C}_0 = \{[0, 1]^d\}$ and $\mathcal{C}_n = \{C_{i_1, \dots, i_n} : i_j \in \{0, \dots, 2^d - 1\}\}$ is a coordinate-plane partition satisfying property (*), with each cell the union of its 2^d children. In Proposition B.2: For $d \geq 1$, the full cube-coordinate construction recursively refines closed coordinate cells by cuts near their coordinate midpoints, perturbing the cuts to obtain $\mathbb{P}(\partial C) = 0$ for each cell. It requires $R(\mathcal{C}_{n+1}) \leq (1 + 2^{-n})R(\mathcal{C}_n)$; its cells need not be the standard dyadic cubes. Here $\mathbb{P}$ is the measure after the proof's change to cube coordinates. In Lemma B.3: When $d \geq 1$, $\mathcal{C}_n = \{\prod_{r=1}^d [a_r 2^{-n}, (a_r + 1)2^{-n}] : a_r \in \{0, \dots, 2^n - 1\}\}$; these cells are closed and may share boundaries.

The following lemma is important in the construction of the Hilbert curve.

Lemma B.3 (Algorithm 3 in [25]). *Given a sequence of partitions of the unit cube $\{\mathcal{C}_n\}$ that has property (*), suppose that, when $d \geq 1$, $\mathcal{C}_n$ is the standard dyadic grid of closed cubes*

$$\mathcal{C}_n = \left\{ \prod_{r=1}^d [a_r 2^{-n}, (a_r + 1)2^{-n}] : a_r \in \{0, \dots, 2^n - 1\} \right\}.$$

There exists an ordering for each partition $\mathcal{C}_n$, i.e.,

$$O_n : \{1, \dots, 2^{nd}\} \longrightarrow \{\overline{i_1 \dots i_n} : i_j = 0, \dots, 2^d - 1\},$$

such that the following two conclusions hold.

1. *If $(j - 1) \cdot 2^d < i \leq j \cdot 2^d$, the first n digits of $O_{n+1}(i)$ are the same as $O_n(j)$.*
2. *Adjacent cubes in the ordering intersect, i.e., $C_{O_n(i)} \cap C_{O_n(i+1)} \neq \emptyset$. Furthermore, $C_{O_n(i)} \cap C_{O_n(i+1)}$ is a $d - 1$ dimensional cube.*

Remark. The cells of $\mathcal{C}_n$ are closed and may share boundaries. Each O_n is a bijection. The assertion about adjacent cubes is quantified over i satisfying $1 \leq i < 2^{nd}$.

Remark (notation). For each positive refinement level n , O_n is an ordering, hence a bijection, from $\{1, \dots, 2^{nd}\}$ onto $\{\overline{i_1 \dots i_n} : i_j \in \{0, \dots, 2^d - 1\}\}$. In Lemma B.3: The chosen sequence preserves parent prefixes: the first n digits of $O_{n+1}(i)$ equal $O_n(j)$ when $(j-1)2^d < i \leq j2^d$; for $1 \leq i < 2^{nd}$, consecutive cells $C_{O_n(i)}$ and $C_{O_n(i+1)}$ share a $(d-1)$ -dimensional face. In Proposition B.2: The proof chooses the sequence $\{O_n\}$ for its recursively constructed cube partitions with parent-prefix consistency: the first n digits of $O_{n+1}(i)$ equal $O_n(j)$ for $(j-1)2^d < i \leq j2^d$. Consecutive cells share a $(d-1)$ -dimensional face for $1 \leq i < 2^{nd}$, and this ordering determines the order of the corresponding mass-weighted parameter intervals.

Proof. For $d = 0$, every partition has its single point as its only cell. Choose the unique ordering; the first conclusion holds and there are no adjacent pairs. For $d \geq 1$, use the recursive Hilbert ordering on the dyadic lattice described in Section 2 of [25]. That construction visits every cell once, visits all the children of a cell consecutively before moving to another cell, and joins successive cells across a coordinate face. Label each child by its position among the 2^d children of its parent, and define $O_n(i)$ to be the resulting address of the i th cell at level n . Visiting every cell once makes O_n a bijection. Completing the children of the j th parent before visiting the next parent places precisely the indices $(j-1)2^d + 1, \dots, j2^d$ in that parent, so their addresses have prefix $O_n(j)$. Successive dyadic lattice cells differ by one unit in exactly one coordinate and agree in the other coordinates. Their closed cubes therefore intersect in the full common face, which has dimension $d-1$ and side length 2^{-n} . This proves both conclusions. $\square$

Proof of Proposition B.2. Without loss of generality, we can assume that $\mathbb{P}$ is supported on $[0, 1]^d$, vanishes at the boundary and $\mathbb{P}(B) \geq C_1 \lambda(B)$ for some constant $C_1 > 0$ and all measurable sets $B \in [0, 1]^d$. Suppose we have proven the conclusion for this case. By ?? 3.5, there exists $f \in \mathcal{H}([0, 1]^d, K) \cap \mathcal{F}_L([0, 1]^d, K)$ such that $f_{\#}^{-1} \mathbb{P}$ satisfies these conditions. There exists continuous function γ such that $f_{\#}^{-1} \mathbb{P} = \gamma_{\#} \lambda$, which implies $\mathbb{P} = (f \circ \gamma)_{\#} \lambda$. In particular, if γ is $1/d$ -Holder continuous, $f \circ \gamma$ is also $1/d$ -Holder continuous because f is Lipschitz continuous.

When $d = 0$, the constant map satisfies the requirement. We focus on the case $d \geq 1$ in the following. The proof is to modify the construction of the Hilbert space-filling curve. By changing the velocity of the curve, the push-forward measure of the Hilbert space-filling curve can simulate a great number of distributions. We use λ to represent the Lebesgue measure $[0, 1]$.

We prove the cube case by first constructing a uniform Hilbert curve and then changing its parameter. Write λ_d for Lebesgue measure on $[0, 1]^d$ and retain λ for Lebesgue measure on $[0, 1]$. The cube probability measure satisfies $\mathbb{P} \geq C_1 \lambda_d$, so $0 < C_1 \leq 1$.

Remark (notation). R is a locally specified shape-ratio functional or fixed positive intervention bound. In Proposition B.2: $R(\mathcal{C}) = \max_{i,k} |a_i^k - b_i^k| / \min_{i,k} |a_i^k - b_i^k|$ for the finite collection $\mathcal{C} = \{C_1, \dots, C_m\}$ with $C_k = \prod_{i=1}^n [a_i^k, b_i^k]$, wherever the denominator is positive. In Corollary 3.8: $R > 0$ is fixed before choosing $\widehat{\mathcal{M}}$ and bounds each continuous intervention value by $\|t_i\|_\infty \leq R$; the implicit approximation constant may depend on R , which remains fixed as the network budgets tend to infinity and $\tau \downarrow 0$.

Step 1: Define partition. First, we construct a sequence of partitions that has the property (*) recursively. Let $\mathcal{C}_0 = \{[0, 1]^d\}$. Construct $\mathcal{C}_{n+1}$ by bisecting every coordinate interval of every cell of $\mathcal{C}_n$ at its midpoint. Label each parent's 2^d children by $i_{n+1} \in \{0, \dots, 2^d - 1\}$ and append this label to the parent's address. Induction gives

$$\mathcal{C}_n = \left\{ \prod_{r=1}^d [a_r 2^{-n}, (a_r + 1) 2^{-n}] : a_r \in \{0, \dots, 2^n - 1\} \right\}.$$

These closed cube collections have the following properties.

1. $\mathcal{C}_n$ is obtained by cutting $[0, 1]^d$ by planes parallel to the coordinate planes.
2. Every parent is the union of its 2^d children, and

$$C_{i_1, \dots, i_n} = \bigcup_{i_{n+1}=0}^{2^d-1} C_{i_1, \dots, i_n, i_{n+1}}, \quad \text{diam}(C_{i_1, \dots, i_n, i_{n+1}}) = \frac{1}{2} \text{diam}(C_{i_1, \dots, i_n}).$$

3. For every $C \in \mathcal{C}_n$, $\lambda_d(\partial C) = 0$ and $\lambda_d(C) = 2^{-nd}$.

The midpoint halves cover each coordinate interval, so their products cover the parent. Each child edge is half the corresponding parent edge, proving the diameter equality. Every level- n edge has length 2^{-n} , giving the displayed volume. Each boundary is a finite union of coordinate faces, each having a singleton in one coordinate and hence zero product Lebesgue measure. Thus these collections satisfy property (*), and every level- n cell has 1-norm diameter $d2^{-n}$.

Step 2: Construct Hilbert Curve. By Lemma B.3, the standard dyadic collections $\mathcal{C}_n$ admit orderings O_n satisfying the following properties.

1. If $(j-1)2^d < i \leq j2^d$, the first n digits of $O_{n+1}(i)$ are the same as $O_n(j)$.
2. Consecutive level- n cubes share a full coordinate face of dimension $d-1$.

Let $\mathcal{I}_0 = \{[0, 1]\}$, we define $\mathcal{I}_n$ recursively. For $n \geq 1$, define

$$I_{O_n(i)} = [(i-1)2^{-nd}, i2^{-nd}], \quad 1 \leq i \leq 2^{nd}, \quad \mathcal{I}_n = \{I_{O_n(i)} : 1 \leq i \leq 2^{nd}\}.$$

Every assigned interval has length $2^{-nd} = \lambda_d(C_{O_n(i)})$. By the prefix conclusion of Lemma B.3, the children of the i th level- n cube have precisely the indices $(i-1)2^d + 1, \dots, i2^d$ at level $n+1$. Their intervals are consecutive, with first left endpoint $(i-1)2^d 2^{-(n+1)d} = (i-1)2^{-nd}$ and last right endpoint $i2^d 2^{-(n+1)d} = i2^{-nd}$. Their union therefore equals $I_{O_n(i)}$. The level-one intervals likewise cover $[0, 1]$. The child lengths sum to $2^d 2^{-(n+1)d} = 2^{-nd}$; the child cube volumes have the same sum, with common faces of zero λ_d -measure. Consequently, for $n \geq 1$ and $1 \leq i \leq 2^{nd}$,

$$|I_{O_n(i)}| = \lambda_d(C_{O_n(i)}) = \sum_{j=(i-1)2^d+1}^{i2^d} \lambda_d(C_{O_{n+1}(j)}) = \sum_{j=(i-1)2^d+1}^{i2^d} |I_{O_{n+1}(j)}|. \quad (\text{B.10})$$

Construct a continuous polygonal curve H_n on these intervals: its initial and final points are the centers of the first and last cubes, and at each internal interval endpoint its value is a point in the common face of the two successive cubes. On each interval interpolate linearly between its assigned endpoint values. Convexity places that segment inside its assigned cube. For a parameter at an interval endpoint, both incident cubes contain the assigned value.

The nesting of the orderings implies that for $m \geq n$, $H_m(t)$ and $H_n(t)$ lie in one common level- n closed cube. At interval endpoints choose the incident cube to the right, except at $t = 1$, where the last cube is chosen. Consequently

$$\sup_{t \in [0, 1]} \|H_m(t) - H_n(t)\| \leq d2^{-n}.$$

The curves converge uniformly to a continuous map $H : [0, 1] \rightarrow [0, 1]^d$, and H maps every assigned level- n interval into its assigned cube.

Let $0 \leq u < v \leq 1$, and choose $n \geq 0$ with $2^{-(n+1)d} < v - u \leq 2^{-nd}$. The two parameters lie in the same or successive level- n intervals. The corresponding cubes coincide or share a face. Their union has 1-norm diameter at most $2d2^{-n}$. Thus

$$\|H(v) - H(u)\| \leq 2d2^{-n} \leq 4d(v - u)^{1/d}.$$

For any continuous scalar function φ on the cube, choose the center $z_{n,i}$ of each level- n cube. Both

$$\int_0^1 \varphi(H(t)) dt \quad \text{and} \quad \int_{[0,1]^d} \varphi(x) d\lambda_d(x)$$

differ from $\sum_i 2^{-nd} \varphi(z_{n,i})$ by at most the modulus of continuity of φ at $d2^{-n}$. Letting $n \rightarrow \infty$ proves $H_{\#}\lambda = \lambda_d$. The compact image of H therefore contains the support of λ_d , which is the entire cube; hence H is surjective.

For each $x \in [0, 1]^d$, define $s(x) = \min\{t : H(t) = x\}$. The fiber is a nonempty compact set. Moreover, $\{x : s(x) \leq a\} = H([0, a])$ is compact for every $a \in [0, 1]$, so s is Borel measurable and $H \circ s$ is the identity. The measure $\mathbb{P} - C_1\lambda_d$ is nonnegative and has mass $1 - C_1$. Define the probability measure on the interval

$$\rho = C_1\lambda + s_{\#}(\mathbb{P} - C_1\lambda_d).$$

It satisfies $\rho \geq C_1\lambda$ and

$$H_{\#}\rho = C_1\lambda_d + (H \circ s)_{\#}(\mathbb{P} - C_1\lambda_d) = \mathbb{P}.$$

This construction allows atoms in ρ and in $\mathbb{P}$.

Put $A(x) = \rho([0, x])$ and define $Q(0) = 0$ and

$$Q(t) = \inf\{x \in [0, 1] : A(x) \geq t\}, \quad 0 < t \leq 1.$$

The domination of ρ gives positive mass to every nonempty open interval, so $Q(1) = 1$. Right continuity of A gives, for $t > 0$, the equivalence $Q(t) \leq x$ if and only if $t \leq A(x)$. Therefore $\lambda\{t : Q(t) \leq x\} = A(x)$, which proves $Q_{\#}\lambda = \rho$.

If $u < v$ and $Q(u) < Q(v)$, take $Q(u) < a < b < Q(v)$. The quantile definition gives $A(a) \geq u$ and $A(b) < v$, also when $u = 0$ or $v = 1$. Hence

$$C_1(b - a) \leq \rho((a, b]) = A(b) - A(a) \leq v - u.$$

Letting $a \downarrow Q(u)$ and $b \uparrow Q(v)$ yields $Q(v) - Q(u) \leq C_1^{-1}(v - u)$. If $Q(u) = Q(v)$ this bound is immediate. Thus Q is Lipschitz and continuous. The cube curve $\zeta = H \circ Q$ satisfies

$$\zeta_{\#}\lambda = H_{\#}\rho = \mathbb{P}, \quad \|\zeta(v) - \zeta(u)\| \leq 4dC_1^{-1/d}|v - u|^{1/d}.$$

For a measure on the original compact set K , apply this construction to $(f^{-1})_{\#}\mathbb{P}$ and take $\gamma = f \circ \zeta$. Then $\gamma_{\#}\lambda = \mathbb{P}$, and the Lipschitz property of f preserves the exponent $1/d$. Since the domination assumption and the homeomorphism imply $\text{supp}(\mathbb{P}) = K$, this curve has the asserted codomain. $\square$

Remark (notation). For $i \in [N_C]$, γ_i represents the pullback of the normalized component law μ_i through g_2^i , wherever $\mathbb{P}(C_i) > 0$. For $i \in [N_C]$ and $j \in [d_i^C]$, $\hat{g}_1^{i,j} : [0, 1] \rightarrow \mathbb{R}$ denotes a coordinate network chosen to approximate $(\gamma_i)_j$ uniformly on $[0, 1]$.

B.5 Proof of results in Section 3.2

Proof of Corollary 3.8. Suppose that structure equations of $\mathcal{M}$ are

$$V_i = \begin{cases} f_i(\text{Pa}(V_i), \mathbf{U}_{V_i}), & V_i \text{ is continuous,} \\ \arg \max_{k \in [n_i]} \{g_k^{V_i} + \log(f_i(\text{Pa}(V_i), \mathbf{U}_{V_i}))_k\}, & V_i \text{ is categorical, } \|f_i\|_1 = 1 \end{cases}, \quad f_i \in \mathcal{F}_L.$$

Put $m_i = d_i^{\text{in}}$, let d_i^{out} be the output dimension of f_i , and put $m = d_{\max}^{\text{in}} > 0$. In each structural input, distinguish the real coordinates, consisting of continuous parents and non-Gumbel latent coordinates, from the categorical parent coordinates. Let a_i have value zero in every real coordinate and value one in every categorical coordinate. Define the fixed constants

$$B_0 = \max(\{1, R, K\} \cup \{n_i : V_i \text{ is categorical}\} \cup \{\|f_i(a_i)\|_{\infty} : V_i \text{ is continuous}\}), \quad H = 1 + 2Lm, \quad B =$$

These constants do not depend on the intervention or on the network budgets.

We first bound all true structural inputs for the interventions in the statement. Work on the probability-one event where every non-Gumbel latent coordinate has absolute value at most K . Order the observed nodes topologically. We claim that the node in position ℓ has infinity norm at most $B_0 H^{\ell}$. An intervened continuous node has norm at most $R \leq B_0$, and a categorical node has value at most its cardinality, which is at most B_0 . For a nonintervened continuous node in position

ℓ , the induction hypothesis bounds all its parents and latent coordinates by $b = B_0 H^{\ell-1}$. Every coordinate of a_i has absolute value at most one, so its actual input x_i satisfies $\|x_i - a_i\| \leq 2m_i b \leq 2mb$. Consequently,

$$\|f_i(x_i)\|_\infty \leq \|f_i(a_i)\|_\infty + L\|x_i - a_i\| \leq B_0 + 2Lmb \leq Hb = B_0 H^\ell.$$

This proves the induction, including its first position. In particular, every true structural input belongs to $Q_i = [-B, B]^{m_i}$, simultaneously for all permitted interventions. For $m_i = 0$, interpret Q_i as the unique point of $\mathbb{R}^0$.

We next define a Lipschitz extension to real values of the categorical input coordinates. A categorical input coordinate with only one category is always replaced by one. For each remaining categorical input coordinate r , with category count $c_r \geq 2$, clip its value x_r to $t_r \in [1, c_r]$, and set

$$k_r = \min\{c_r - 1, \lfloor t_r \rfloor\}, \quad \theta_r = t_r - k_r \in [0, 1].$$

Let s_i be the number of these remaining categorical coordinates. For $\beta \in \{0, 1\}^{s_i}$, let x^β retain the real coordinates of x , replace coordinate r by $k_r + \beta_r$, and replace every one-category coordinate by one. With $w_{r,0} = 1 - \theta_r$ and $w_{r,1} = \theta_r$, define

$$f_i^\dagger(x) = \sum_{\beta \in \{0,1\}^{s_i}} \left(\prod_{r=1}^{s_i} w_{r,\beta_r} \right) f_i(x^\beta).$$

When $s_i = 0$, the sum has one term with weight one. Thus this definition includes functions with no categorical parents. At valid categorical labels it agrees with f_i .

This interpolation preserves the L -Lipschitz bound in the paper's 1-norm. Changing one real coordinate leaves the interpolation weights fixed, so the difference is a convex combination of differences bounded in 1-norm by L times the coordinate change. With all other coordinates fixed, on each interval between consecutive labels a categorical coordinate gives a linear interpolation between two vectors whose 1-norm difference is at most L . Its slope therefore has 1-norm at most L ; outside the label interval the function is constant in that coordinate. Splitting any coordinate change at its finitely many intervening labels proves the same bound for an arbitrary change of that coordinate. Changing all coordinates one at a time and adding the bounds gives

$$\|f_i^\dagger(x) - f_i^\dagger(y)\| \leq L\|x - y\|.$$

For a categorical output, the interpolation weights are nonnegative and sum to one, so $f_i^\dagger(x) \in \Delta_{n_i}$ for every real input x .

Let $\Pi_i : \mathbb{R}^{m_i} \rightarrow Q_i$ be coordinatewise projection, explicitly

$$(\Pi_i x)_r = -B + \max\{x_r + B, 0\} - \max\{x_r - B, 0\},$$

and define $\bar{f}_i = f_i^\dagger \circ \Pi_i$. Clipping each coordinate cannot increase its distance from another clipped coordinate; summing proves $\|\Pi_i x - \Pi_i y\| \leq \|x - y\|$. Hence $\bar{f}_i$ is globally L -Lipschitz. Since $a_i \in Q_i$, for every $z \in Q_i$ and every continuous output,

$$\|f_i^\dagger(z)\|_\infty \leq \|f_i(a_i)\|_\infty + L\|z - a_i\| \leq B_0 + 2Lm_i B \leq HB.$$

Thus $\bar{f}_i$ is globally bounded by HB for continuous outputs; categorical outputs remain probability vectors.

Use the same latent laws and Gumbel noises to form the model with structural functions $\bar{f}_i$. Couple it with the original model using identical draws. Intervened nodes agree by their assigned values. If the parents of a nonintervened node agree, their common original input belongs to Q_i and has valid categorical coordinates. Projection fixes this input and interpolation agrees there with f_i , so the structural outputs agree; for a categorical node use the identical Gumbel draws and the convention $\log 0 = -\infty$. Induction in topological order proves equality of the two interventional laws for every intervention in the statement. The extended true model satisfies the boundedness assumption with the fixed enlarged constant HB , because its nonintervened continuous outputs are globally bounded by HB , its categorical labels are at most B_0 , its assigned treatment coordinates are at most B_0 , and its non-Gumbel latent coordinates are at most B_0 .

We now choose the neural approximants in place of the exact-minimizer construction. Reserve $c = 4$ hidden layers of the structural budget for projection, categorical floors, and padding. For each $m_i > 0$, apply [52, Theorem 2], part b), to every coordinate of $f_i^\dagger$ on Q_i , after the affine rescaling $z \mapsto 2Bz - B\mathbf{1}$ from $[0, 1]^{m_i}$. The rescaled coordinate functions have Lipschitz moduli bounded by a fixed constant times their argument. The source gives scalar width $2m_i + 10$ and error at most a fixed constant times the modulus at a constant times the number of weights raised to $-2/m_i$. At fixed width that number is proportional to hidden depth. Parallelizing the d_i^{out} coordinates therefore gives, for every sufficiently large L_0 , a network q_i of width at most $d_i^{\text{out}}(2m_i + 10)$ and hidden depth at most $L_0 - c$ such that

$$\sup_{z \in Q_i} \|q_i(z) - f_i^\dagger(z)\|_\infty \leq C_i(L_0 - c)^{-2/m_i}.$$

The constant C_i is fixed independently of L_0 . For $L_0 \geq 2c$, choose a fixed $A_i \geq \max\{1, 2^{2/m_i} C_i\}$ and set

$$\epsilon_i = A_i L_0^{-2/m_i}.$$

Then choose one of the networks just supplied, so that its uniform error on Q_i is at most ϵ_i . Define the structural networks by

$$\widehat{f}_i(x) = \begin{cases} q_i(\Pi_i x), & V_i \text{ is continuous,} \\ (\max\{\epsilon_i, (q_i(\Pi_i x))_k\})_{k=1}^{n_i}, & V_i \text{ is categorical.} \end{cases}$$

For $m_i = 0$, instead set $\epsilon_i = 0$ and let $\widehat{f}_i = f_i$ be the exact constant affine network. These definitions specify $\widehat{f}_i$ without any choice of a minimizer.

Projection uses one hidden layer with at most $2m_i$ units, and each categorical floor is $z \mapsto \epsilon_i + \max\{z - \epsilon_i, 0\}$, using one further hidden layer. Affine maps between these layers and the network q_i are absorbed into the adjacent affine transformations. If signed coordinates must be carried through a hidden layer, use $z = \max\{z, 0\} - \max\{-z, 0\}$ with two units per coordinate. Thus the reserved four layers suffice, with width at most $4m_i + 2d_i^{\text{out}}(2m_i + 10) + 2d_i^{\text{out}}$. For $m_i \geq 1$ and $d_i^{\text{out}} \geq 1$, this is at most $30m_i d_i^{\text{out}} \leq 64d_{\max}^{\text{in}} d_{\max}^{\text{out}}$, because $2m_i + 10 \leq 12m_i$. A constant map with $m_i = 0$ uses at most $2d_i^{\text{out}}$ units per hidden layer by carrying its positive and negative parts, which also fits this universal width since $d_{\max}^{\text{in}} \geq 1$. A zero-dimensional output needs no units. Zero padding of units and insertion of identity layers on nonnegative hidden outputs give the structural architecture and hidden depth L_0 in the statement.

The continuous global approximation error is at most ϵ_i , because $\Pi_i x \in Q_i$ for every input x . For categorical outputs, if $p \geq 0$ and $\epsilon \geq 0$, then

$$|\max\{\epsilon, z\} - p| \leq \max\{\epsilon, |z - p|\}.$$

Indeed, for $z \geq \epsilon$ the left side is $|z - p|$; for $z < \epsilon$ and $p \leq \epsilon$ it is at most ϵ ; and for $z < \epsilon < p$ it is $p - \epsilon \leq p - z$. Apply this inequality coordinatewise at $\Pi_i x$ to obtain

$$\sup_{x \in \mathbb{R}^{m_i}} \|\widehat{f}_i(x) - \bar{f}_i(x)\|_\infty \leq \epsilon_i.$$

For $m_i > 0$, every categorical coordinate is at least $\epsilon_i > 0$. For $m_i = 0$, the exact constant vector is a probability vector. In both cases its 1-norm is strictly positive. The global error bound holds also when generated latent or parent inputs lie outside Q_i .

For a categorical variable, at an arbitrary input x , put $p = \bar{f}_i(x)$, $q = \widehat{f}_i(x)$, $a = \|q\|_1$, and $S = \|q - p\|_\infty$. The construction gives $a > 0$ and $S \leq \epsilon_i$. Since p is a probability vector and q is nonnegative,

$$|a - 1| \leq \sum_{j=1}^{n_i} |q_j - p_j| \leq n_i S.$$

For sufficiently large L_0 , simultaneously for all nodes with $m_i > 0$, we have $n_i \epsilon_i \leq 1/2$, and hence $a \geq 1/2$. Since $\|p\|_\infty \leq 1$,

$$\begin{aligned} \left\| \frac{q}{a} - p \right\|_\infty &\leq \frac{\|q - p\|_\infty + |1 - a| \|p\|_\infty}{a} \\ &\leq 2(n_i + 1)S \leq 2(n_i + 1)\epsilon_i. \end{aligned}$$

For $m_i = 0$, normalization is exact because $q = p$ and $a = 1$. For $m_i > 0$, the inequalities $m_i \leq d_{\max}^{\text{in}}$ and $L_0 \geq 1$ imply $L_0^{-2/m_i} \leq L_0^{-2/d_{\max}^{\text{in}}}$. Taking the supremum over every real structural input therefore gives

$$\left\| \bar{f}_i - \frac{\widehat{f}_i}{\|\widehat{f}_i\|_1} \right\|_\infty \leq O(L_0^{-2/d_{\max}^{\text{in}}}).$$

The continuous structural error has the same order without normalization. These bounds use the fixed cardinalities n_i and do not require $n_i \leq K$. For each non-Gumbel latent variable U_j with positive dimension d_j^U , apply the proposition entitled “Deep distribution approximation” to its law. Choose the resulting generators with independent uniform and Gumbel inputs Z_j and put $\widehat{U}_j = \widehat{g}_j(Z_j)$. Each component cube dimension is at most the ambient dimension d_j^U . For budgets at least one, the component bounds therefore give

$$W(P(U_j), P(\widehat{U}_j)) \leq O\left(L_1^{-2/d_j^U} + L_2^{-2/d_j^U} + (\tau - \tau \log \tau)\right).$$

For a component of positive cube dimension $q \leq d_j^U$, the scalar first-block networks have width 12, and the scalar second-block networks have width $2q + 10 \leq 12q$. Parallelization gives widths

$12q$ and $12d_j^U q$. Coordinate clipping needs at most $2q$ or $2d_j^U$ units, and carrying signed values requires at most twice the corresponding width; reserving four layers therefore gives widths at most $64q$ and $64d_j^U q$. A singleton component uses its exact constant blocks; carrying a signed constant second-block output needs at most $2d_j^U$ units. Since $d_{\max}^U \geq 1$, every component fits widths $64d_{\max}^U$ and $64(d_{\max}^U)^2$. These are the per-component widths in the separable architecture, so the number of components does not multiply either width budget. Zero padding and identity layers on nonnegative hidden outputs give the architecture in the statement. The numerical factor 64 is independent of the model, R , dimensions, and cardinalities; only the error constants and minimum depth budgets may depend on the fixed model. A latent vector of dimension zero is the unique point of $\mathbb{R}^0$ and is represented exactly, with error zero. Independence of the original latent variables and independence of the chosen generator inputs allow application of Lemma E.3, yielding

$$\begin{aligned} W(P(U_1, \dots, U_{n_U}), P(\widehat{U}_1, \dots, \widehat{U}_{n_U})) &\leq \sum_{j=1}^{n_U} W(P(U_j), P(\widehat{U}_j)) \\ &\leq O\left(L_1^{-2/d_{\max}^U} + L_2^{-2/d_{\max}^U} + (\tau - \tau \log \tau)\right). \end{aligned}$$

Indeed, for $d_j^U > 0$, the inequalities $d_j^U \leq d_{\max}^U$ and $L_r \geq 1$ give $L_r^{-2/d_j^U} \leq L_r^{-2/d_{\max}^U}$ for $r = 1, 2$, and there are only finitely many latent variables. Identical categorical Gumbel noises may be adjoined to both latent vectors without changing this bound: couple these noises identically for one inequality and project onto the non-Gumbel coordinates for the reverse inequality. Let $\widehat{\mathcal{M}}$ be the casual model with the following structure equations

$$\widehat{V}_i = \begin{cases} \widehat{f}_i\left(\text{Pa}(\widehat{V}_i), (\widehat{g}_j(Z_{C_j}))_{U_{C_j} \in U_{V_i}}\right), & V_i \text{ is continuous,} \\ \arg \max_{k \in [n_i]} \left\{ g_k + \log \left(\widehat{f}_i\left(\text{Pa}(\widehat{V}_i), (\widehat{g}_j(Z_{C_j}))_{U_{C_j} \in U_{V_i}}\right) \right)_k \right\}, & V_i \text{ is categorical.} \end{cases}$$

Replacing a categorical propensity vector $\widehat{f}_i(x)$ by $\widehat{f}_i(x)/\|\widehat{f}_i(x)\|_1$ subtracts the same finite constant $\log \|\widehat{f}_i(x)\|_1$ from every finite log propensity. It therefore preserves the categorical structural equation, with $\log 0 = -\infty$ for zero coordinates. Apply Theorem 3.3 to the globally L -Lipschitz true functions $\bar{f}_i$, the continuous approximate functions $\widehat{f}_i$, and the normalized approximate categorical functions. The extended true model has boundedness constant HB and, by the coupling induction in

this proof, the same interventional laws as $\mathcal{M}^*$ for every intervention in the statement. Consequently,

$$\begin{aligned}
& W \left(P^{\mathcal{M}^*}(\mathbf{V}(\mathbf{t})), P^{\widehat{\mathcal{M}}}(\mathbf{V}(\mathbf{t})) \right) \\
& \leq C \left(\sum_{V_i \text{ continuous}} \|\bar{f}_i - \widehat{f}_i\|_\infty + \sum_{V_i \text{ categorical}} \left\| \bar{f}_i - \frac{\widehat{f}_i}{\|\widehat{f}_i\|_1} \right\|_\infty + W(P(U_1, \dots, U_{n_U}), P(\widehat{U}_1, \dots, \widehat{U}_{n_U})) \right) \\
& \leq O \left(L_0^{-2/d_{\max}^{\text{in}}} + L_1^{-2/d_{\max}^U} + L_2^{-2/d_{\max}^U} + (\tau - \tau \log \tau) \right).
\end{aligned}$$

For the second inequality, every term of the first sum has the continuous structural error bound proved here, every term of the second sum has the normalized categorical error bound proved here, and the latent term has the bound obtained using Lemma E.3; the two sums contain only finitely many nodes. The comparison constant C depends only on the fixed model, its dimensions and cardinalities, and R . The cubes, extensions, and networks were chosen independently of the intervention within the prescribed region. This proves the estimate uniformly throughout that region. $\square$

By [52, Theorem 2], part b), applied with $p = 2/d_i^{\text{in}}$ to each output coordinate of f_i , which is Lipschitz continuous on the input cube by ?? 3.2: there are narrow fully-connected ReLU networks of constant width $2d_i^{\text{in}} + 10$ and depth L_0 whose sup-norm error is at most $a\omega(cW^{-2/d_i^{\text{in}}})$, where W , the number of weights, grows linearly in L_0 at fixed width; stacking the d_i^{out} coordinates gives, if V_i is continuous, $\|\widehat{f}_i - f_i\|_\infty \leq O(L_0^{-2/d_i^{\text{in}}}) \leq O(L_0^{-2/d_{\max}^{\text{in}}})$.

Theorem 2 b) of [52] applied to each output coordinate, whose modulus of continuity satisfies $\omega(r) \leq L_f r$; at fixed width the number of weights W is proportional to the depth L_0 , so $a\omega(cW^{-2/d_i^{\text{in}}}) = O(L_0^{-2/d_i^{\text{in}}})$.

C Proof of Consistency

C.1 Inconsistent Counterexample (Proposition 4.1)

Remark (notation). d_j^U , for $j \in [n_U]$, is the Euclidean coordinate dimension of the indexed latent variable U_j , excluding the separately introduced categorical Gumbel noises.

Proposition C.1. *There exists a constant $c > 0$ and an SCM $\mathcal{M}^*$ satisfying Assumptions 2.3-3.5 with a backdoor causal graph such that there is no unobserved confounding in $\mathcal{M}^*$. For any*

$\epsilon > 0$, there exists an SCM $\mathcal{M}_\epsilon$ with the same causal graph such that $W(P^{\mathcal{M}^*}(\mathbf{V}), P^{\mathcal{M}_\epsilon}(\mathbf{V})) \leq \epsilon$ and $|ATE_{\mathcal{M}^*} - ATE_{\mathcal{M}_\epsilon}| > c$. Here, for each non-Gumbel latent law $\mathbb{P}$ with connected support components C_i , Assumption 3.5 is required of each normalized component law $\mathbb{P}|_{C_i}/\mathbb{P}(C_i)$, with $K = C_i$.

Additional condition. In the categorical structural equations, $\arg \max$ selects the smallest maximizing label when there is a tie. This is a measurable completion on ties; under the stipulated independent standard Gumbel noises and nonnegative propensity weights of positive total, it leaves every fixed-intervention law unchanged.

Remark. The second Bernoulli law in the construction has $P(U_2 = 0) = 2/3$.

Proof. Let $\mathbf{V} = \{X, T, Y\}$ be the covariate, treatment and outcome respectively. T is a binary variable. Let structure equations of causal model $\mathcal{M}_\delta$ be

$$\begin{aligned} X &= \begin{cases} -1 & , \text{w.p. } 1/2 \\ 0 & , \text{w.p. } 1/4 \\ \delta & , \text{w.p. } 1/4, \end{cases} \\ P(T = 1|X = x) &= \begin{cases} 1/2 & , x = -1 \text{ or } 0, \\ 3/4 & , x = \delta, \end{cases} \\ Y &= \begin{cases} T + U_y & , X < 0 \\ X/\delta + U_y & , X \geq 0 \end{cases} \end{aligned}$$

where all latent variables are independent and U_y is mean-zero noise. The distribution of $\mathcal{M}_\delta$ is

$$\begin{aligned} P_{X,T,Y}(0, 1, y) &= p_{U_y}(y)/8, P_{X,T,Y}(0, 0, y) = p_{U_y}(y)/8 \\ P_{X,T,Y}(\delta, 1, y) &= 3p_{U_y}(y - 1)/16, P_{X,T,Y}(\delta, 0, y) = p_{U_y}(y - 1)/16. \\ P_{X,T,Y}(-1, 1, y) &= p_{U_y}(y - 1)/4, P_{X,T,Y}(-1, 0, y) = p_{U_y}(y)/4. \end{aligned}$$

As $\delta \rightarrow 0$, distribution of $\mathcal{M}_\delta \xrightarrow{d} \mathcal{M}^*$, where structure equations of $\mathcal{M}^*$ are

$$P(U_1 = 1) = 3/5, P(U_1 = 0) = 2/5, P(U_2 = 1) = 1/3, P(U_2 = 0) = 2/3,$$

$$X = \begin{cases} -1 & , \text{w.p. } 1/2, \\ 0 & , \text{w.p. } 1/2, \end{cases}$$

$$P(T = 1|X = x) = \begin{cases} 1/2 & , x = -1, \\ 5/8 & , x = 0, \end{cases}$$

$$Y = \begin{cases} T + U_y & , X = -1, \\ U_1 + U_y & , X = 0, T = 1, \\ U_2 + U_y & , X = 0, T = 0, \end{cases}$$

where U_1, U_2, U_y are independent. The distribution of $\mathcal{M}^*$ is

$$P_{X,T,Y}(0, 1, y) = p_{U_y}(y)/8 + 3p_{U_y}(y-1)/16,$$

$$P_{X,T,Y}(0, 0, y) = p_{U_y}(y)/8 + p_{U_y}(y-1)/16,$$

$$P_{X,T,Y}(-1, 1, y) = p_{U_y}(y-1)/4,$$

$$P_{X,T,Y}(-1, 0, y) = p_{U_y}(y)/4.$$

Choose $U_y = 0$ almost surely. Generate X by $X = U_X$, where $P(U_X = -1) = P(U_X = 0) = 1/2$, independently of U_1, U_2 and the categorical Gumbel variables. Every non-Gumbel latent law, including any vector formed by merging latents affecting the same observed variable, now has finite support. For such a law $\mathbb{P}$, write its support components as $C_i = \{a_i\}$ and put $\mu_i = \mathbb{P}|_{C_i}/\mathbb{P}(C_i)$. The map $f : [0, 1]^0 \rightarrow C_i$ sending the unique point to a_i is a Lipschitz homeomorphism. If λ_0 denotes the probability measure on $[0, 1]^0$, then for every measurable $B \subset [0, 1]^0$,

$$\mu_i(f(B)) = \lambda_0(B).$$

Thus each normalized component law satisfies Assumption 3.5 with $d = 0$ and $C_f = 1$; its boundary requirement applies only when $d > 0$. These singleton components also verify Assumption 3.4. All displayed structural inputs range over finite sets. Their structural and normalized propensity maps have Lipschitz extensions: form the finite Cartesian grid of their input coordinate values, retain the prescribed values at the specified inputs, and assign bounded values at the remaining vertices, using probability vectors for propensity maps. Interpolate multilinearly on each grid

cell and clamp each input coordinate to the grid's range. This preserves the prescribed values and the probability simplex. Each cell has finite coordinate slopes; taking their maximum and subdividing a line segment at cell boundaries gives a finite global Lipschitz constant. Choose L_f to bound these finitely many constants. The resulting observed variables and non-Gumbel latent variables are bounded, so a finite K exists. Independence and categorical generation hold by construction. Hence $\mathcal{M}^*$ satisfies Assumptions 2.3-3.5. Some calculation gives $W(P^{\mathcal{M}^*}(\mathbf{V}), P^{\mathcal{M}_\delta}(\mathbf{V})) \leq \delta/2$. It is easy to calculate the ATE of the two models.

$$\text{ATE}_{\mathcal{M}^*} = 19/30, \text{ATE}_{\mathcal{M}_\delta} = 1/2.$$

□

This example implies that even as the Wasserstein distance between $P^{\mathcal{M}^*}(\mathbf{V}), P^{\mathcal{M}_\delta}(\mathbf{V})$ converges to zero, their ATEs do not change. As mentioned in the main body, this problem is caused by the violation of Lipschitz continuity assumption for $\mathcal{M}_\delta$. Note that in $\mathcal{M}_\delta$, $\mathbb{E}[Y|X, T] = X/\delta$. As $\delta \rightarrow 0$, the Lipschitz constant explodes.

This problem may also arise in (4.1). If the distribution ball $B_n = \{\widehat{\mathcal{M}} : W(P^{\widehat{\mathcal{M}}}, P_n^{\mathcal{M}^*}) \leq \alpha_n\}$ includes a small ball around the true distribution $S_\epsilon = \{\widehat{\mathcal{M}} : W(P^{\widehat{\mathcal{M}}}, P^{\mathcal{M}^*}) \leq \epsilon\} \subset B_n$ and the NCM is expressive enough to approximate all the SCMs in the S_ϵ , this example tells us the confidence interval may not shrink to one point as sample size increases to infinity even if the model is identifiable.

C.2 Proof of Theorem 4.2

To prove Theorem 4.2, we will need the following proposition.

Proposition C.2. *Given a metric space (M, d) and sets $\Theta_1 \subset \Theta_2 \cdots \subset \Theta_n \subset \cdots \subset \Theta_\infty \subset M$, where $\Theta_\infty = \overline{\bigcup_{n=1}^\infty \Theta_n}$, positive sequences $\{\epsilon_n\}_{n \in \mathbb{N}}, \{\delta_n\}_{n \in \mathbb{N}}$ such that $\lim_{n \rightarrow \infty} \epsilon_n = \lim_{n \rightarrow \infty} \tau_n = \lim_{n \rightarrow \infty} \delta_n = 0$ and continuous functions $f, g_n : \Theta_\infty \rightarrow \mathbb{R}$, suppose that*

1. Θ_∞ is compact.

2. g_n satisfies

$$\|g_n(\theta_1) - g_n(\theta_2)\| \leq L_g d(\theta_1, \theta_2) + \tau_n, \quad \forall \theta_1, \theta_2 \in \Theta \quad (\text{C.1})$$

and $\sup_{\theta \in \Theta_\infty} \|g_n(\theta) - g(\theta)\| \leq \delta_n, g_n \geq 0$.

3. There exists a compact subset $\tilde{\Theta}_\infty \subset \Theta_\infty$ such that for any $\theta_0 \in \tilde{\Theta}_\infty$, there exists $\theta \in \Theta_n$ such that $d(\theta, \theta_0) \leq \epsilon_n$.

Consider the following optimization problems:

$$\begin{aligned} \min_{\theta \in \Theta_n} f(\theta), \\ \text{s.t. } g_n(\theta) \leq L_g \epsilon_n + \delta_n, \end{aligned} \quad (\text{C.2})$$

$$\begin{aligned} \min_{\theta \in \Theta_\infty} f(\theta), \\ \text{s.t. } g(\theta) = 0. \end{aligned} \quad (\text{C.3})$$

and

$$\begin{aligned} \min_{\theta \in \tilde{\Theta}_\infty} f(\theta), \\ \text{s.t. } g(\theta) = 0. \end{aligned} \quad (\text{C.4})$$

We assume that the feasible region of (C.3) and (C.4) are nonempty. Let $f_n^*, f^*, \tilde{f}^*$ be the minimal of (C.2), (C.3) and (C.4) respectively. Then, $[\liminf_{k \rightarrow \infty} f_n^*, \limsup_{k \rightarrow \infty} f_n^*] \subset [f^*, \tilde{f}^*]$.

Additional condition. For every $n \in \mathbb{N}$, the feasible set of (C.2) contains a minimizer of f .

Remark. In (C.1), $\Theta = \Theta_\infty$. Equation C.1 holds for every $n \in \mathbb{N}$ and every $\theta_1, \theta_2 \in \Theta_\infty$, with the same finite constant L_g . There exists a compact $\tilde{\Theta}_\infty \subseteq \Theta_\infty$ such that $\forall n \in \mathbb{N} \forall \theta_0 \in \tilde{\Theta}_\infty \exists \theta \in \Theta_n$, $d(\theta, \theta_0) \leq \epsilon_n$.

Remark (notation). $\Theta_n \subset M$ is the parameter set indexed by $n \in \mathbb{N}$. In Proposition C.2: $\Theta_1 \subset \Theta_2 \subset \dots \subset \Theta_n \subset \dots$, and every $\theta_0 \in \tilde{\Theta}_\infty$ has a point $\theta \in \Theta_n$ with $d(\theta, \theta_0) \leq \epsilon_n$. In Theorem 4.2: $\Theta_n = \hat{\mathcal{F}}_{V_1, n} \times \dots \times \hat{\mathcal{F}}_{V_{n_V}, n} \times \mathcal{P}_n$. For continuous V_i , $\hat{\mathcal{F}}_{V_i, n} = \{\hat{f}_i \in \mathcal{F}_{0, n} : \sup_x \|\hat{f}_i(x)\|_1 \leq K\}$, with the supremum over the full structural input domain. For categorical V_i , $\hat{\mathcal{F}}_{V_i, n}$ consists of $x \mapsto q(x)/\|q(x)\|_1$ with $q \in \mathcal{F}_{0, n}$ satisfying $q(x) \in [0, \infty)^{n_i}$, $\|q(x)\|_1 > 0$ for every structural input x , and $\text{Lip}(x \mapsto q(x)/\|q(x)\|_1) \leq \hat{L}_f$. Here $\mathcal{P}_n = \{P(\mathbf{U}) : \mathbf{U} \text{ is the latent distribution of } \widehat{\mathcal{M}} \in \text{NCM}_{\mathcal{G}}(\mathcal{F}_{0, n}, \mathcal{F}_{1, n})\}$, identified with its tuple of independent latent marginal laws. This product records population model parameters; the empirical optimization ranges over their network representations, including distinct representations inducing identical latent population laws. $\Theta_\infty = \overline{\bigcup_{n=1}^\infty \Theta_n}$, with closure taken in the ambient metric space (M, d) .

In Proposition C.2: Θ_∞ is compact and contains the nested sets Θ_n ; $\{\theta \in \Theta_\infty : g(\theta) = 0\}$ is nonempty. In Theorem 4.2: The closure is that of the neural-model parameter sets in the metric $d(\theta, \theta') = \sum_{k=1}^{n_V} \|f_k - f'_k\|_\infty + \sum_{k=1}^{n_U} W(P(U_k), P(U'_k))$ on the product space of structural functions and latent probability measures. $\tilde{\Theta}_\infty$ is a compact subset of Θ_∞ approximated by Θ_n : for every $\theta_0 \in \tilde{\Theta}_\infty$ and every n , some $\theta \in \Theta_n$ satisfies $d(\theta, \theta_0) \leq \epsilon_n$. In Proposition C.2: $\{\theta \in \tilde{\Theta}_\infty : g(\theta) = 0\}$ is nonempty. δ_n denotes a deterministic error tolerance indexed by n and tending to zero. In Proposition C.2: $\delta_n > 0$ for $n \in \mathbb{N}$, and $\sup_{\theta \in \Theta_\infty} |g_n(\theta) - g(\theta)| \leq \delta_n$. In a result cited in the proof of Theorem 4.2: For sufficiently large n , the local observational latent deviation tolerance is $\delta_n = C_1 n^{-1/\max\{d_{\max}^U, 2\}} \log^2 n$, with fixed $C_1 > 0$ chosen sufficiently large; this is distinct from the abstract uniform constraint-error tolerance. $\{g_n\}_{n \in \mathbb{N}}$ is a sequence of real-valued constraint functions on Θ_∞ . In Proposition C.2: Each g_n is continuous and nonnegative, with $\sup_{\theta \in \Theta_\infty} |g_n(\theta) - g(\theta)| \leq \delta_n$, and the approximate Lipschitz bound has common coefficient L_g and additive tolerance τ_n . In a result cited in the proof of Theorem 4.2: The notation is the empirical constraint $g_n(\theta) = S_{\lambda_n}(P_n^{\theta^*}(\mathbf{V}), P_{m_n}^\theta(\mathbf{V}))$, compared with the population constraint $g(\theta) = W(P^{\theta^*}(\mathbf{V}), P^\theta(\mathbf{V}))$.

Proof of Proposition C.2. By compactness of Θ_∞ and $\tilde{\Theta}_\infty$, the minimizers of (C.3) and (C.4) are achievable insider these two sets. Let $\theta^* \in \Theta_\infty, \tilde{\theta}^* \in \tilde{\Theta}_\infty$ be the minimizer of (C.3) and (C.4). We first prove that $\limsup_{n \rightarrow \infty} f_n^* \leq \tilde{f}^*$. Note that by (C.1), the limiting function g is L_g -Lipschitz. By condition 3, there exist $\theta_n \in \Theta_n$ such that $d(\theta_n, \tilde{\theta}^*) \leq \epsilon_n$. Note that

$$\begin{aligned} g_n(\theta_n) &\leq |g(\theta_n) - g_n(\theta_n)| + |g(\theta_n) - g(\theta^*)| + |g(\theta^*)| \\ &\leq \delta_n + L_g d(\theta_n, \theta^*) \leq \delta_n + L_g \epsilon_n. \end{aligned}$$

Therefore, θ_n is a feasible point of (C.2). We have $\limsup_{n \rightarrow \infty} f_n^* \leq \limsup_{n \rightarrow \infty} f(\theta_n) = f(\tilde{\theta}^*) = \tilde{f}^*$.

Remark (notation). L_g is a finite nonnegative Lipschitz bound. In Proposition C.2: L_g is independent of n and is the coefficient in $|g_n(\theta_1) - g_n(\theta_2)| \leq L_g d(\theta_1, \theta_2) + \tau_n$; the limiting function g is L_g -Lipschitz. In Lemma C.3: In the proof, $L_g = 2K \max\{L, 1\}$ bounds the Lipschitz constant of $g_0(t, y) = (\hat{f}(t) - f(t))y$ on $\{(t, y) : \|(t, y)\|_\infty \leq K\}$. θ^* denotes a distinguished parameter point whose role is specified by the local optimization or causal-model parameterization. In Proposition C.2: $\theta^* \in \Theta_\infty$ is a minimizer of (C.3), satisfying $g(\theta^*) = 0$ and $f(\theta^*) = f^*$; it is distinct in notation and role from $\tilde{\theta}^*$, the minimizer over $\tilde{\Theta}_\infty$. In Theorem 4.2: θ^* is the canonical

parameter tuple of the underlying SCM, containing its structural functions and non-Gumbel latent marginal laws; $\tilde{\Theta}_\infty = \{\theta^*\}$. In Corollary 4.4: θ^* is the canonical parameter tuple representing the underlying SCM $\mathcal{M}^*$, containing its structural functions and non-Gumbel latent population marginal laws.

Next, we argue that $\liminf_{n \rightarrow \infty} f_n^* \geq f^*$. If this equation does not hold, there exists $\epsilon > 0$ and subsequence $\{f_{n_k}^*\}_{k \in \mathbb{N}}$ such that $f_{n_k}^* < f^* - \epsilon, \forall k \in \mathbb{N}$. By compactness of Θ_∞ , for each k , there exists a subsequence of $\theta_{n_k}^* \in \Theta_\infty$ such that $\theta_{n_k}^*$ satisfies constraint of (C.2) and $f_{n_k}^* = f(\theta_{n_k}^*)$. By compactness, $\{\theta_{n_k}^*\}_{k \in \mathbb{N}}$ has a converging subsequence. Without loss of generality, we may assume that $\{\theta_{n_k}^*\}_{k \in \mathbb{N}}$ converges to $\hat{\theta}^* \in \Theta_\infty$. Since g_k converge uniformly to g ,

$$\lim_{k \rightarrow \infty} g_k(\theta_{n_k}^*) \leq \lim_{k \rightarrow \infty} |g_k(\theta_{n_k}^*) - g(\theta_{n_k}^*)| + g(\theta_{n_k}^*) = g(\hat{\theta}^*) \leq \lim_{k \rightarrow \infty} \alpha_k = 0.$$

$\hat{\theta}^*$ is a feasible point of Equation (C.3). Since f is continuous on Θ ,

$$f(\hat{\theta}^*) = \lim_{k \rightarrow \infty} f(\theta_{n_k}^*) \leq f^* - \epsilon.$$

which leads to contradiction. □

Proof of Theorem 4.2. We begin by defining a proper metric space. By ?? 3.2, all random variables are bounded. Suppose that $\max_{i,j} \{\|V_i\|_\infty, \|U_j\|_\infty\} \leq K$. A canonical causal model $\mathcal{M} \in \mathcal{M}(\mathcal{G}, \mathcal{F}, \mathbf{U})$ is decided by

$$\theta_{\mathcal{M}} = (f_1, \dots, f_{n_V}, P(U_1), \dots, P(U_{n_U})),$$

where f_i are functions in the structure equations (3.1) and U_j are uniform parts of the latent variables. We denote $\mathcal{M}^\theta$ to be the SCM represented by θ , the underlying SCM be $\mathcal{M}^{\theta^*}$ and P^θ to be the distribution of $\mathcal{M}^\theta$. We consider the space

$$M = \mathcal{F}_{V_1} \times \dots \times \mathcal{F}_{V_{n_V}} \times \prod_{j=1}^{n_U} \mathcal{P}([-K, K]^{d_j^U}),$$

where

$$\mathcal{F}_{V_i} = \begin{cases} \mathcal{F}_{\hat{L}_f}^K([-K, K]^{d_{i,\text{in}}^V}, [-K, K]^{d_{i,\text{out}}^V}), & V_i \text{ continuous,} \\ \{p : [-K, K]^{d_{i,\text{in}}^V} \rightarrow \Delta_{n_i} : \text{Lip}(p) \leq \hat{L}_f\}, & V_i \text{ categorical.} \end{cases}$$

Here $d_{i,\text{in}}^V, d_{i,\text{out}}^V$ are the input and output dimensions of f_i , $\mathcal{F}_{\widehat{L}_f}^K$ has uniform output 1-norm bound K and common Lipschitz bound $\widehat{L}_f$, and $\mathcal{P}(A)$ denotes the probability measures on A . For $\theta = (f_1, \dots, f_{n_V}, P(U_1), \dots, P(U_{n_U}))$, $\theta' = (f'_1, \dots, f'_{n_V}, P(U'_1), \dots, P(U'_{n_U}))$, we define a metric on M

$$d(\theta, \theta') = \sum_{k=1}^{n_V} \|f_k - f'_k\|_\infty + \sum_{k=1}^{n_U} W(P(U_k), P(U'_k)).$$

We derive the population continuity bound in the fixed dimensions of this theorem. Couple corresponding non-Gumbel latent laws optimally and take the product of these couplings, using their independence. The expected joint latent 1-norm discrepancy is the sum of their marginal Wasserstein distances. Generate the observed nodes in a topological order. At a continuous node the expected output discrepancy is at most the structural output 1-norm approximation error plus $\widehat{L}_f$ times the expected input discrepancy. At a categorical node with conditional propensity vectors p, p' , a maximal coupling has disagreement probability $\|p - p'\|_1/2$; its expected label distance is at most $(n_i - 1)\|p - p'\|_1/2$. Such a coupling assigns common mass $\min\{p_k, p'_k\}$ to label k and distributes the remaining masses on the disjoint residual supports. It preserves each model's conditional categorical law. If e is the sum of the expected latent and previously generated observed discrepancies, adjoining a continuous node increases it to at most $(1 + \widehat{L}_f)e + d_i\|f_i - f'_i\|_\infty$; adjoining a categorical node increases it to at most $(1 + (n_i - 1)\widehat{L}_f/2)e + n_i(n_i - 1)\|f_i - f'_i\|_\infty/2$. Choose fixed constants $A \geq 1$ and B bounding these finitely many coefficients. Iterating $e_{k+1} \leq Ae_k + Bd(\theta, \theta')$ gives

$$e_k \leq A^k e_0 + Bd(\theta, \theta') \sum_{r=0}^{k-1} A^r.$$

Indeed, the formula holds for $k = 0$, and multiplying its right side by A and adding $Bd(\theta, \theta')$ gives the formula for $k + 1$. Identical intervention values contribute zero discrepancy and remove their generation steps. Projection to the observed coordinates therefore proves

$$W(P^\theta(\mathbf{V}(\mathbf{t})), P^{\theta'}(\mathbf{V}(\mathbf{t}))) \leq Cd(\theta, \theta')$$

uniformly over the intervention values, with C depending on the fixed graph, dimensions, cardinalities and regularity bounds. The triangle inequality consequently makes $g(\theta) = W(P^{\theta^*}(\mathbf{V}), P^\theta(\mathbf{V}))$ Lipschitz in d ; denote a Lipschitz constant by C_g . Now, we define Θ_n . Let

$$\mathcal{P}_n = \left\{ P(\mathbf{U}) : \mathbf{U} \text{ is the latent distribution of } \widehat{\mathcal{M}} \in \text{NCM}_{\mathcal{G}}(\mathcal{F}_{0,n}, \mathcal{F}_{1,n}) \right\}.$$

Thus $\mathcal{P}_n$ consists of the latent population laws obtained by pushing forward the independent uniform and generator Gumbel inputs through the current positive-temperature generator networks. We denote $\Theta_n = \widehat{\mathcal{F}}_{V_1,n} \times \cdots \times \widehat{\mathcal{F}}_{V_{n_V},n} \times \mathcal{P}_n$, where

$$\widehat{\mathcal{F}}_{V_i,n} = \begin{cases} \{f \in \mathcal{F}_{0,n} : \sup_x \|f(x)\|_1 \leq K\}, & V_i \text{ continuous,} \\ \{q/\|q\|_1 : q \in \mathcal{F}_{0,n}, q(x) \in [0, \infty)^{n_i}, \|q(x)\|_1 > 0 \forall x, \text{Lip}(q/\|q\|_1) \leq \widehat{L}_f\}, & V_i \text{ categorical.} \end{cases}$$

The norms and Lipschitz bounds are evaluated on the structural input domains, and $\mathcal{F}_{0,n}$ is defined in Theorem 4.2. Thus $\Theta_n \subset M$. Note that by construction, latent variables are independent and $\mathcal{P}_n$ can be decomposed into direct product $\mathcal{P}_{n,1} \times \cdots \times \mathcal{P}_{n,n_U}$. Let $\mathcal{A}_n$ be the set of restricted structural and generator network representations at sample size n , and let $\theta_\eta \in \Theta_n$ be the population parameter of a representation $\eta \in \mathcal{A}_n$. Let $\xi_1, \dots, \xi_{m_n}$ be the common independent exogenous draws, including the categorical Gumbel coordinates, and write $H_\eta(\xi_l)$ for the resulting complete observed vector. Define

$$\Theta_\infty = \overline{\bigcup_n \Theta_n}, \quad P_{m_n}^\eta = \frac{1}{m_n} \sum_{l=1}^{m_n} \delta_{H_\eta(\xi_l)}, \quad h_n(\eta) = S_{\lambda_n}(P_n^{\theta^*}, P_{m_n}^\eta),$$

$$f(\theta) = \mathbb{E}_{t \sim \mu_T} \mathbb{E}_{\mathcal{M}^\theta} [F(V_1(t), \dots, V_{n_V}(t))].$$

The empirical constraint is a function of the representation η , whereas the population objective and $g(\theta)$ are functions of θ . The optimization minimizes $f(\theta_\eta)$ over representations satisfying $h_n(\eta) \leq \alpha_n$.

For fixed n , pad the finitely many network architectures to their stated widths and encode each representation by its finitely many real weights and mixture probabilities. Finite products of Euclidean spaces and simplices are standard Borel spaces. Affine maps, ReLU, clipping and finite composition have jointly continuous evaluation in their weights and inputs. At the fixed temperature $\tau_n > 0$, mixture weights can be written

$$\frac{p_j^{1/\tau_n} \exp(G_j/\tau_n)}{\sum_k p_k^{1/\tau_n} \exp(G_k/\tau_n)}, \quad p \in \Delta,$$

whose denominator is positive even when some $p_j = 0$; hence generator evaluation is jointly Borel.

The restrictions defining $\mathcal{A}_n$ are Borel restrictions on these parameters. To check this, choose a countable dense set in each continuous structural input domain. Continuity makes the output bound, coordinatewise nonnegativity and every Lipschitz inequality equivalent to their inequalities on that dense set or its Cartesian square. Each such inequality is Borel in the weights. For a nonnegative categorical raw network q , positivity on a compact input box D_0 is equivalent to membership in

$$\bigcup_{l=1}^{\infty} \bigcap_{x \in Q_0} \{\eta : \|q_\eta(x)\|_1 \geq 1/l\},$$

where Q_0 is countable and dense in D_0 ; continuity and attainment of a minimum prove both implications. Use the intersection of these conditions over a countable compact exhaustion when the full input domain is unbounded. On this Borel parameter set $q/\|q\|_1$ is jointly Borel and continuous in the input, so its Lipschitz restriction is again a countable family of Borel inequalities. The same dense-input tests apply to the raw Lipschitz restrictions before clipping. Thus $\mathcal{A}_n$ is standard Borel. These tests use dense input sets only.

The exogenous input space is a finite product of uniform cubes and real Gumbel coordinates, hence standard Borel. For a categorical propensity vector p , put $r_i = \exp(-g_i)/p_i$ when $p_i > 0$ and $r_i = \infty$ when $p_i = 0$. These extended-real functions are Borel. The smallest minimizing label is i exactly when $r_i < r_j$ for every $j < i$ and $r_i \leq r_j$ for every $j > i$; this is a finite intersection of Borel comparisons. Induction through the finite topological order therefore proves joint Borel evaluation of $H_\eta(\xi)$ and of its interventional version $H_\eta(t, \xi)$. The function F is bounded on the bounded observed range, so integration over the fixed exogenous law and μ_T makes $\eta \mapsto f(\theta_\eta)$ Borel. Finite empirical optimal transport and entropic transport values are continuous in their cost matrices: changing every cost by at most a changes every transport objective, and hence its infimum, by at most a . Therefore the empirical Sinkhorn constraint $h_n(\eta)$ is jointly Borel in the representation and data. For every real a , the event $\{F_n < a\}$ is the projection of the Borel set defined by $h_n(\eta) \leq \alpha_n$ and $f(\theta_\eta) < a$, so it is analytic. The empirical optimal value and its limit events are measurable in the completed sample laws. If L_F is a Lipschitz constant of F , the population coupling bound gives

$$|f(\theta) - f(\theta')| \leq L_F \mathbb{E}_{t \sim \mu_T} W(P^\theta(\mathbf{V}(t)), P^{\theta'}(\mathbf{V}(t))) \leq L_F C d(\theta, \theta').$$

Thus f is continuous. We establish compactness, a sampling estimate for the current representations, and approximation of the true population parameter.

1. Every continuous structural class has a common uniform output 1-norm bound K and Lipschitz constant $\widehat{L}_f$ on a compact input domain. Every categorical structural class takes values in the compact simplex Δ_{n_i} and has the same common Lipschitz bound. Arzela–Ascoli makes each class relatively compact in the uniform norm. These classes are closed: for continuous maps a uniform limit satisfies $\|f(x)\|_1 = \lim_r \|f_r(x)\|_1 \leq K$ at every input x , uniform limits preserve the closed simplex constraint for categorical maps, and passing to the limit in $\|f_r(x) - f_r(y)\|_1 \leq \widehat{L}_f \|x - y\|_1$ preserves the Lipschitz constraint. Hence every $\mathcal{F}_{V_i}$ is compact. Probability measures on each compact latent cube form a compact space in the weak topology by Prokhorov’s theorem. On a compact cube Wasserstein distance metrizes this topology. The finite product M is consequently compact in d , and its closed subset Θ_∞ is compact.
2. We bound the complete observed simulation error, retaining paired observations and all categorical noise. Put $H_n = 1 + L_{1,n} + L_{2,n}$. The finite component counts of the specified separable architectures are fixed. For $0 < \rho < 1/2$, round the joint non-Gumbel latent vector to a grid of 1-norm mesh at most ρ in its D_U -dimensional cube. There are at most $C\rho^{-D_U}$ grid points. For each grid point and each possible categorical prefix in a topological order, evaluate the structural equations with that prefix fixed. Record the continuous outputs and categorical propensity vectors in a table, and round them within ρ , using a finite simplex grid for propensities. There are finitely many prefixes, independent of n . The tables have a finite set $\mathcal{T}_\rho$ of possible values with

$$\log |\mathcal{T}_\rho| \leq C\rho^{-D_U} \log(2 + \rho^{-1}).$$

Indeed, each entry lies in a fixed bounded finite-dimensional box, so a grid has at most $(C/\rho)^C$ choices, and the number of entries is at most $C\rho^{-D_U}$. With a categorical prefix fixed, induction through the continuous structural equations bounds their change under a latent perturbation of size ρ by $C\rho$: each new difference is at most $\widehat{L}_f$ times the sum of previous and latent differences. The categorical propensities have the same bound. Thus, as long as the selected categorical prefixes agree, the original and table outputs differ by at most $C\rho$, and the corresponding propensities differ coordinatewise by at most $a = C\rho$.

The table simulation uses the original categorical Gumbel draws. Write $E_k = \exp(-g_k)$;

these are independent unit-rate exponential variables, and a propensity vector p selects $\arg \min_k E_k/p_k$, with $E_k/0 = \infty$. For a table propensity vector $q \in \Delta_N$, call label i robust if $q_i > a$ and

$$\frac{E_i}{q_i - a} < \min_{j \neq i} \frac{E_j}{q_j + a}.$$

A robust label wins for every propensity vector whose coordinates differ from q by at most a . The robust-label events are disjoint. For $q_i > a$, integrating the density of E_i and the survival probabilities of the other exponentials gives

$$\mathbb{P}(i \text{ robust}) = \int_0^\infty \exp \left(-s \left(1 + \frac{\sum_{j \neq i} (q_j + a)}{q_i - a} \right) \right) ds = \frac{q_i - a}{q_i - a + \sum_{j \neq i} (q_j + a)}.$$

For $q_i \leq a$ this probability is zero. Consequently,

$$\mathbb{P}(i \text{ robust}) \geq \frac{(q_i - a)_+}{1 + Na}.$$

Summing and using $\sum_i (q_i - a)_+ \geq 1 - Na$ bounds the probability of no robust label by $2Na/(1 + Na)$ when $Na < 1$, and the trivial bound $1 \leq 2Na$ covers $Na \geq 1$. For each table and generator representation let b be the indicator that a nonrobust choice occurs at any table prefix. Independence of the categorical noise from the generated latent vector and a union bound over the finitely many prefixes give $\mathbb{E}b \leq C\rho$. If $b = 0$, induction over the categorical nodes forces the original and table prefixes to agree. Consequently, writing $\tilde{H}$ for the table observed map and B_V for a common bound on the observed 1-norm diameter,

$$\|H_\eta(\xi) - \tilde{H}(\xi)\|_1 \leq C\rho + B_V b(\xi).$$

This is a pointwise bound, including zero propensities.

We bound the complexity of finite-type indicators and ambiguity indicators. A table simulation has type (u, c) , where u is its latent grid index and c its complete categorical label vector; its observed output is a fixed vector for each type. There are at most $J = C\rho^{-D_V}$ types. Include the indicators of membership in every subset of these types, and the ambiguity indicators b , in a scalar class $\mathcal{B}_{n,\rho}$.

Here is an explicit standard Borel parameterization of this entire indicator class. Let Γ_n be the finite-dimensional generator weight space with its mixture simplex factors, and let generator parameters range over all of Γ_n . This supplies an envelope for every restricted generator used in the optimization. A parameter consists of $\gamma \in \Gamma_n$, a table in the finite set $\mathcal{T}_\rho$, and a finite tag specifying either a subset of the at most J types or the ambiguity indicator. This is a finite union of products of standard Borel spaces. Generator evaluation is jointly Borel by the positive-denominator mixture formula. Round its clipped latent output using fixed half-open grid cells, with the last cell closed at the endpoint; the grid index is a finite-valued Borel function. For each fixed table and grid index, the table simulation uses finitely many categorical race comparisons. In the notation $r_i = \exp(-g_i)/q_i$, with $r_i = \infty$ at $q_i = 0$, the smallest-label rule is the finite system $r_i < r_j$ for $j < i$ and $r_i \leq r_j$ for $j > i$, so its selected label is Borel. The robust-label events are Borel too: when $q_i > a$ all their denominators are positive, and when $q_i \leq a$ the event is empty. Their finite unions and complements make b jointly Borel. Type-subset membership is a finite union of these Borel events. Thus $\mathcal{B}_{n,\rho}$ has jointly Borel evaluation on a standard Borel parameter space, including zero propensities, grid boundaries and all ties. When using the ambient Euclidean input space in Lemma E.6, clip the uniform input coordinates to their cubes; this changes no sample value and preserves joint Borel evaluation. The parameter envelope has precisely the same architecture and finite tags as those counted in the growth bound. For each continuous scalar dual test h , its normalized square-root transform is jointly Borel by composition. Consequently the structural condition of Lemma E.6 holds for these families, without approximation of selector parameters.

On M_0 fixed exogenous inputs, the latent grid indices are determined by at most CM_0/ρ scalar generator comparisons. The ReLU blocks have $P \leq CH_n$ parameters and at most CH_n layers. Within a fixed activation pattern, their outputs are polynomials in the parameters of degree at most CH_n . For a positive-temperature mixture write its weights as $a_j R_j / \sum_k a_k R_k$, where $a_j = p_j^{1/\tau_n} \geq 0$, $\sum_j a_j > 0$, and $R_j = \exp(G_j/\tau_n)$ is fixed on each input. Treat the a_j as additional free parameters. Clearing the positive denominator in a grid comparison leaves a polynomial of degree at most $CH_n + 1$. Truncation adds only fixed-degree comparisons. The polynomial sign bound in Lemma 17 of [6], applied successively to the activation comparisons in each layer and then to these grid comparisons, gives at most

$$(C(M_0 + H_n)H_n/\rho)^{CH_n^2}$$

latent grid patterns: at each layer the sign bound contributes at most $(C(M_0 + H_n)H_n/\rho)^{CP}$, and there are at most CH_n layers. For a fixed table and grid pattern, the categorical race comparisons and every table output on these inputs are determined. The subgraph growth of $\mathcal{B}_{n,\rho}$ is therefore at most this number times $|\mathcal{T}_\rho|2^J$: a subset of the types has 2^J choices, and an ambiguity indicator needs no additional choices. Solving the shattering inequality

$$2^{M_0} \leq |\mathcal{T}_\rho|2^J (C(M_0 + H_n)H_n/\rho)^{CH_n^2}$$

bounds their pseudo-dimensions by

$$v_n(\rho) \leq C (\rho^{-D_U} \log(2 + \rho^{-1}) + H_n^2 \log(2 + H_n/\rho)) .$$

To derive this bound, put $A = \log |\mathcal{T}_\rho| + J \log 2$ and $B = CH_n^2$. The shattering inequality gives $M_0 \log 2 \leq A + B \log(C(M_0 + H_n)H_n/\rho)$. If $M_0 \leq 4A/\log 2$, the first term suffices. Otherwise, absorbing A and solving the remaining logarithmic inequality gives $M_0 \leq CB \log(2 + BH_n/\rho) + CB \log(2 + A/B)$; this follows from $\log x \leq x/2$ after rescaling M_0/B . The bound on A makes this at most the displayed expression.

Put

$$r_n = m_n^{-1/(D+2)} \log m_n + H_n m_n^{-1/2} \log m_n,$$

and take $\rho = r_n$ for sufficiently large n . The depth schedule implies $H_n = O(m_n^{d_{\max}^U/(2d_{\max}^U+4)})$, so $r_n \rightarrow 0$. Apply Lemma E.6 to $\mathcal{B}_{n,r_n}$ with scalar output. Its functions take values in $\{0, 1\}$, its pseudo-dimension is at most $v_n(r_n)$, and Wasserstein distance between its Bernoulli output laws equals the absolute difference of their means. Choose the lemma's tolerances $\delta = r_n$ and $\epsilon = A_0 r_n$. Its exponent is bounded above by

$$C \left((1 + r_n^{-1}) \log(2 + r_n^{-1}) + r_n^{-D_U} \log^2(2 + r_n^{-1}) + H_n^2 \log(2 + H_n/r_n) \log(2 + r_n^{-1}) \right) - cA_0^2 m_n r_n^2.$$

Here $\log(2 + H_n/r_n) = O(\log m_n)$ and

$$m_n r_n^2 \geq m_n^{D/(D+2)} \log^2 m_n + H_n^2 \log^2 m_n,$$

whereas $r_n^{-D_U} \leq m_n^{D_U/(D+2)} (\log m_n)^{-D_U}$ and $r_n^{-1} \leq m_n^{1/(D+2)} / \log m_n$. Since $D \geq D_U + 1$, a sufficiently large fixed A_0 makes the failure probability at most $O(m_n^{-3})$. On the complementary event every type-subset probability and every ambiguity-indicator mean differs from its empirical counterpart by at most $C r_n$. The supremum over type subsets is the total variation distance between the population and empirical type laws, so that distance is at most $C r_n$. Couple their common type masses identically and their residual masses arbitrarily. Because table outputs have diameter at most B_V , their Wasserstein distance is at most $B_V C r_n$. For each original representation, coupling original and table outputs on the same draw bounds their population error by $C\rho + B_V \mathbb{E}b$ and their empirical error by $C\rho + B_V m_n^{-1} \sum_l b(\xi_l)$. The ambiguity estimate bounds both by $C r_n$. The triangle inequality therefore gives

$$\sup_{\eta \in \mathcal{A}_n} W(P^{\theta_\eta}(\mathbf{V}), P_{m_n}^\eta(\mathbf{V})) \leq C r_n$$

with probability at least $1 - O(m_n^{-3})$. This retains the positive-temperature generator, paired simulation observations and every categorical noise coordinate.

For the observational sample let C denote the complete categorical label vector, which has a fixed finite set of possible values. Conditional on $C = c$, all continuous observations equal a deterministic map $\Phi_c(U_1, \dots, U_{n_U})$: fix the categorical values to c and evaluate the continuous structural equations in topological order. Induction through their Lipschitz bounds makes Φ_c Lipschitz with a constant uniform over these finitely many c . The conditional joint latent laws $\mu_c = P((U_1, \dots, U_{n_U}) \mid C = c)$ are supported on the bounded D_U -dimensional latent cube; they need not be products. Write $p_c = P(C = c)$ and N_c for the observational sample counts. For $p_c = 0$, $N_c = 0$ almost surely. For every $p_c > 0$, a binomial concentration bound gives $N_c \geq np_c/2$ except on an exponentially small event. Conditional on the label vector of the sample, the latents of the observations carrying label c are independent with law μ_c .

When $D_U \geq 1$, Theorem 2 of [17], applied to each conditional joint latent empirical measure of size N_c , bounds its Wasserstein error by $C_c n^{-1/\max\{2, D_U\}} \log^2 n$ with failure probability

$O(n^{-4})$ whenever $N_c \geq np_c/2$. Its small-tolerance exponents are constant multiples of $N_c x^2$ for $D_U = 1$, $N_c(x/\log(2 + 1/x))^2$ for $D_U = 2$, and $N_c x^{D_U}$ for $D_U > 2$. At the indicated x these grow at least as $\log^4 n$, $\log^2 n$ and $\log^{2D_U} n$, respectively. Bounded support supplies the exponential-moment condition; norm equivalence changes only constants. Conditioning on the counts preserves the bound, since it holds uniformly for $np_c/2 \leq N_c \leq n$. Lipschitz push-forward by Φ_c bounds each conditional observed Wasserstein error by the same order. If $D_U = 0$, each conditional observed law is a point mass and its conditional sampling error is zero. Binomial concentration also bounds $\sum_c |N_c/n - p_c|$ by $C\sqrt{\log n/n}$ with probability at least $1 - O(n^{-4})$. Couple common categorical-component masses using the conditional couplings and unmatched masses arbitrarily within the observed support. This gives

$$W(P_n^{\theta^*}(\mathbf{V}), P^{\theta^*}(\mathbf{V})) \leq \sum_c \min\{p_c, N_c/n\} W((\Phi_c)_\# \mu_c, (\Phi_c)_\# (\mu_c)_{N_c}) + \frac{B_V}{2} \sum_c |N_c/n - p_c| \leq t_n,$$

$$t_n = C_1 n^{-1/\max\{2, D_U\}} \log^2 n,$$

with probability at least $1 - O(n^{-3})$, after a union bound over the finitely many categories. Terms with $N_c = 0$ have zero common mass and are omitted. Finally Lemma E.5 and the Wasserstein triangle inequality give, on the combined event,

$$\sup_{\eta \in \mathcal{A}_n} |h_n(\eta) - g(\theta_\eta)| \leq 2(1 + \log(nm_n))\lambda_n + Cr_n + t_n \leq s_n.$$

This event has probability at least $1 - O(n^{-3} + m_n^{-3})$. Since $m_n = \Omega(n)$, the failure probabilities are summable; Borel–Cantelli gives the estimate eventually almost surely. It applies to the current representations.

3. We construct structural approximants satisfying both output and Lipschitz restrictions. Let $f : [-K, K]^m \rightarrow \mathbb{R}^{m'}$ have Lipschitz constant L_f , with $m \geq 1$. Divide each coordinate into R equal intervals of length $h = 2K/R$, and triangulate each cell into simplices with successive vertices $v_0, v_1 = v_0 + he_{\pi(1)}, \dots, v_m = v_{m-1} + he_{\pi(m)}$, where π runs over the coordinate permutations. Let $I_h f$ be affine on each simplex and equal f at its vertices. Its derivative column in direction $e_{\pi(j)}$ is $(f(v_j) - f(v_{j-1}))/h$, of 1-norm at most L_f . Splitting a

line segment at simplex boundaries and summing the derivative bound proves $\text{Lip}(I_h f) \leq L_f$; segments in boundaries follow by continuity. If a_j are the barycentric coordinates of x , then

$$\|I_h f(x) - f(x)\|_1 \leq \sum_{j=0}^m a_j \|f(v_j) - f(x)\|_1 \leq L_f m h.$$

For categorical f , the interpolated values are convex combinations in Δ_{n_i} , so they are nonnegative and sum exactly to one. For continuous f , the additional uniform output bound on the true structural function gives

$$\|I_h f(x)\|_1 \leq \sum_j a_j \|f(v_j)\|_1 \leq K.$$

Thus each continuous approximant satisfies the continuous network restriction.

The nodal basis functions are translates and rescalings of

$$\phi(z) = \max \{0, \min (\{1 + z_k, 1 - z_k : k \in [m]\} \cup \{1 + z_k - z_l : k, l \in [m]\})\}.$$

On each simplex in the star of the central vertex, the active affine expression is its barycentric coordinate: the inequalities defining the simplex order its coordinates, and the minimum selects the difference between the two coordinates bordering that vertex. It equals one at the central vertex and zero at the other vertices; outside the star the minimum is nonpositive. Therefore

$$I_h f(x) = \sum_v f(v) \phi((x - v)/h)$$

on the input cube. Each minimum is computed by a binary tree using $\min(a, b) = a - \max\{a - b, 0\}$. There are $O(m^2)$ affine inputs, so each term has width $O(m^2)$ and depth $O(1 + \log m)$. Parallelizing the terms and summing their outputs gives width $O(R^m)$ in fixed dimensions. Coordinatewise input clipping to the cube is a contraction and makes the network globally defined with the same Lipschitz and output bounds. Choosing R proportional to $W_{0,n}^{1/m}$ gives error $O(W_{0,n}^{-1/m})$. When $m = 0$, represent the constant function exactly. For a categorical function take $q = I_h f$; its normalization equals itself, so both Lipschitz restrictions hold. The output clipping in $\mathcal{F}_{0,n}$ acts identically on these bounded approximants.

We obtain the latent approximants from the inherited component assumptions. For each normalized support component measure μ_i , let $\psi_i : [0, 1]^{q_i} \rightarrow C_i$ be its Lipschitz chart. For $q_i > 0$, Proposition B.2 applied to the pullback measure supplies a curve $\gamma_i : [0, 1] \rightarrow [0, 1]^{q_i}$ of Holder exponent $1/q_i$ with that pullback law. Theorem 2(b) of [52] applied to its coordinates gives first-block uniform error $O(L_{1,n}^{-2/q_i})$; applied to the Lipschitz coordinates of ψ_i it gives second-block error $O(L_{2,n}^{-2/q_i})$. Parallelize the coordinate networks and clip the first outputs to their input cubes. Their widths fit the stated fixed-width budgets. Lemma E.1, including its fixed output-dimension factor, bounds the component Wasserstein error by $O(L_{1,n}^{-2/q_i} + L_{2,n}^{-2/q_i})$. For $q_i = 0$, represent the singleton component by its constant value with zero error. With component probabilities p_i , Lemma E.2 bounds the hard selected population mixture error by the weighted sum of these component errors. The positive-temperature construction uses independent Gumbel-softmax weights; the hard mixture is its zero-temperature comparison law. Clip component outputs to a fixed box containing their supports. For outputs v_i in that box,

$$\left\| \sum_i (X_i^{\tau_n} - X_i^0) v_i \right\|_1 \leq \left(\max_i \|v_i\|_1 \right) \sum_i |X_i^{\tau_n} - X_i^0| \leq C \|X^{\tau_n} - X^0\|_1.$$

Couple these weights using identical Gumbel draws and retain identical independent component outputs. The coupling estimate in Lemma E.4 bounds the additional population error by $O(\tau_n(1 - \log \tau_n))$ for sufficiently large n . The triangle inequality proves, for each non-Gumbel latent U_j , a generator $g_{j,n}$ with

$$W(P(U_j), P(g_{j,n}(Z_j))) \leq O\left(L_{1,n}^{-2/d_{\max}^U} + L_{2,n}^{-2/d_{\max}^U} + \tau_n(1 - \log \tau_n)\right).$$

Coordinatewise output truncation T_K cannot increase this error: $T_K U_j = U_j$ and $\|U_j - T_K g_{j,n}(Z_j)\|_1 \leq \|U_j - g_{j,n}(Z_j)\|_1$ in every coupling. Choose the generators independently. Summing their marginal errors and the structural errors in d gives a representation $\eta_n \in \mathcal{A}_n$ with

$$a_n = d(\theta^*, \theta_{\eta_n}) \leq O\left(W_{0,n}^{-1/d_{\max}^{\text{in}}} + L_{1,n}^{-2/d_{\max}^U} + L_{2,n}^{-2/d_{\max}^U} + \tau_n(1 - \log \tau_n)\right) \rightarrow 0.$$

Choose the constant in ϵ_n large enough that $C_g a_n \leq \epsilon_n$.

Work on the probability-one event where the current-class sampling estimate holds eventually. Since $g(\theta^*) = 0$, the constructed representations satisfy

$$h_n(\eta_n) \leq g(\theta_{\eta_n}) + s_n \leq C_g a_n + s_n \leq \epsilon_n + s_n = \alpha_n.$$

Thus the restricted optimization is eventually feasible. Its objective satisfies $f(\theta_{\eta_n}) \rightarrow f(\theta^*) = F_*$, whence $\limsup_n F_n \leq F_*$. Choose a feasible representation $\bar{\eta}_n$ with $F_n \leq f(\theta_{\bar{\eta}_n}) \leq F_n + 1/n$. The bounded objective on compact M makes F_n eventually bounded. Along a subsequence attaining its lower limit, compactness gives a further subsequence on which $\theta_{\bar{\eta}_n} \rightarrow \bar{\theta}$. Feasibility and the current-class bound imply

$$0 \leq g(\theta_{\bar{\eta}_n}) \leq h_n(\bar{\eta}_n) + s_n \leq \alpha_n + s_n \rightarrow 0.$$

Continuity yields $g(\bar{\theta}) = 0$. Its continuous structural functions satisfy $\sup_x \|f_i(x)\|_1 \leq K$, all structural functions have Lipschitz bound $\hat{L}_f$, its categorical functions are simplex-valued, and its latent marginal laws define an independent bounded canonical latent distribution. Since categorical variables occur only when $K \geq 1$, their normalized propensity vectors also satisfy the output 1-norm bound. Hence the limit is an admissible observationally equivalent model in (4.2), so $f(\bar{\theta}) \geq \underline{F}^{\hat{L}_f}$. Continuity and $|f(\theta_{\bar{\eta}_n}) - F_n| \leq 1/n$ give $\liminf_n F_n = f(\bar{\theta})$. This proves

$$[\liminf_n F_n, \limsup_n F_n] \subset [\underline{F}^{\hat{L}_f}, F_*]$$

almost surely. On the finite-sample event used in the sampling proof, every feasible representation also satisfies

$$W(P^{\theta^*}(\mathbf{V}), P^{\theta_\eta}(\mathbf{V})) \leq h_n(\eta) + s_n \leq \alpha_n + s_n \leq 2\alpha_n.$$

This event has probability at least $1 - O(n^{-3} + m_n^{-3})$ and supplies the complete-observation error and feasibility estimates used in Corollary 4.4 for the same restricted optimizer and revised radius. $\square$

C.3 Proof of Proposition 4.3

Lemma C.3. Let $(\hat{T}, \hat{Y}) \sim \mu, (T, Y) \sim \nu$ and suppose that $f(t) = \mathbb{E}_\nu[Y|T]$, $\hat{f}(t) = \mathbb{E}_\mu[\hat{Y}|\hat{T}]$ are L -Lipschitz continuous and $|f(t)| \leq K, |\hat{f}(t)| \leq K$, then we have

$$\int (f(t) - \hat{f}(t))^2 d\nu(dt) \leq C_W W(\mu, \nu),$$

where $C_W = 4LK + 2K \max\{L, 1\}$.

Additional condition. The laws μ and ν of $(\hat{T}, \hat{Y})$ and (T, Y) are both supported on $[-K, K]^2$, with the same $K > 0$ as in the regression bounds.

Remark. The functions $f, \hat{f} : [-K, K] \rightarrow \mathbb{R}$ are L -Lipschitz regression versions satisfying $f(T) = \mathbb{E}_\nu[Y | T]$ almost surely under ν and $\hat{f}(\hat{T}) = \mathbb{E}_\mu[\hat{Y} | \hat{T}]$ almost surely under μ .

Remark (notation). $\hat{T}$ is the real-valued treatment coordinate of $(\hat{T}, \hat{Y}) \sim \mu$, and $\hat{f}(t) = \mathbb{E}_\mu[\hat{Y} | \hat{T} = t]$. In Lemma C.3: The proof treats $(\hat{T}, \hat{Y})$ as supported in $\{(t, y) : \|(t, y)\|_\infty \leq K\}$.

Proof of Lemma C.3. By the duality formulation of Wasserstein-1 distance, we have

$$W(\mu, \nu) = \sup_{g \in \text{Lip}(1)} \mathbb{E}_\mu[g(\hat{T}, \hat{Y})] - \mathbb{E}_\nu[g(T, Y)].$$

Let $g_0(t, y) = (\hat{f}(t) - f(t))y$, we verify g_0 is a Lipschitz continuous function in $\{(t, y) : \|(t, y)\|_\infty \leq K\}$.

$$\begin{aligned} |g_0(t_1, y_1) - g_0(t_2, y_2)| &\leq |g_0(t_1, y_1) - g_0(t_1, y_2)| + |g_0(t_1, y_2) - g_0(t_2, y_2)| \\ &= |(\hat{f}(t_1) - f(t_1))(y_1 - y_2)| + |y_2(\hat{f}(t_1) - f(t_1) - (\hat{f}(t_2) - f(t_2)))| \\ &\leq 2K|y_1 - y_2| + K(|\hat{f}(t_1) - \hat{f}(t_2)| + |f(t_1) - f(t_2)|) \\ &\leq 2K|y_1 - y_2| + 2KL|t_1 - t_2|. \end{aligned}$$

Let $L_g = 2K \max\{L, 1\}$, we have proven that g_0 is L_g -Lipschitz continuous in $\{(t, y) : \|(t, y)\|_\infty \leq K\}$. Thus,

$$\begin{aligned} W(\mu, \nu) &\geq \frac{1}{L_g} (\mathbb{E}_\mu[g_0(\hat{T}, \hat{Y})] - \mathbb{E}_\nu[g_0(T, Y)]) \\ &= \frac{1}{L_g} (\mathbb{E}_\mu[(\hat{f}(\hat{T}) - f(\hat{T}))\mathbb{E}[\hat{Y}|\hat{T} = t]] - \mathbb{E}_\nu[(\hat{f}(T) - f(T))\mathbb{E}[Y|T = t]]) \\ &= \frac{1}{L_g} (\mathbb{E}_\mu[(\hat{f}(\hat{T}) - f(\hat{T}))\hat{f}(\hat{T})] - \mathbb{E}_\nu[(\hat{f}(T) - f(T))f(T)]). \end{aligned} \tag{C.5}$$

Now, let $h(t) = (\widehat{f}(t) - f(t))\widehat{f}(t)$. Following the same argument, it can be proven that $h(t)$ is Lipschitz continuous with Lipschitz constant being $L_h = 4LK$. Hence,

$$\mathbb{E}_\mu[h(\widehat{T})] \geq \mathbb{E}_\nu[h(T)] - L_h W(\mu, \nu).$$

Plug into (C.5), and we get

$$\begin{aligned} (L_g + L_h)W(\mu, \nu) &\geq \mathbb{E}_\nu[(\widehat{f}(T) - f(T))\widehat{f}(T) - (\widehat{f}(T) - f(T))f(T)] \\ &= \mathbb{E}_\nu[(\widehat{f}(T) - f(T))^2]. \end{aligned}$$

□

Proof of Proposition 4.3. If Y is not a descendant of T , its structural equation and all the equations of its ancestors are unaffected by the intervention, so $\mathbb{E}_\mathcal{M}[Y(t)] = \mathbb{E}_\mathcal{M}[Y]$ and $\mathbb{E}_{\widehat{\mathcal{M}}}[\widehat{Y}(t)] = \mathbb{E}_{\widehat{\mathcal{M}}}[\widehat{Y}]$. For any coupling of the observational laws, the absolute difference of the outcome means is at most the expected absolute difference of the outcome coordinates, which is at most the expected 1-norm distance of the observed vectors. Taking the infimum over couplings gives

$$\begin{aligned} \int_{t_1}^{t_2} (\mathbb{E}_\mathcal{M}[Y(t)] - \mathbb{E}_{\widehat{\mathcal{M}}}[\widehat{Y}(t)])^2 P(dt) &= P([t_1, t_2]) (\mathbb{E}_\mathcal{M}[Y] - \mathbb{E}_{\widehat{\mathcal{M}}}[\widehat{Y}])^2 \\ &\leq P([t_1, t_2]) W^2(\mu, \nu) \\ &\leq \frac{2C_W}{\delta} W(\mu, \nu) + 2L_y^2 P([t_1, t_2]) W^2(\mu, \nu), \end{aligned}$$

since $L_y \geq 1$. Now, suppose that Y is a descendant of T . Let $X = \text{Pa}(T)$, and let A be the vector of all observed nodes other than T that are not descendants of T . In particular, X is a subvector of A . Every observed parent of a node in A also belongs to A : a parent equal to T , or descended from T , would make that node a descendant of T . By the canonical representation and the absence of hidden confounding, the noises belonging to distinct observed nodes are independent. Thus A is generated without the noise of T or any descendant of T , and the descendant noises are independent of (A, T) . Conditional on $A = a$, the equation of T uses only $X = x$ and its own independent noise. Consequently,

$$\nu(dt|A = a) = \nu(dt|X = x).$$

For each $\mathcal{N} \in \{\mathcal{M}, \widehat{\mathcal{M}}\}$, fix $A = a$ and $T = t$ and generate the descendants of T in topological order using their independent noises with their laws in $\mathcal{N}$. Denote the resulting mean of the outcome

by $g_{\mathcal{N}}(a, t)$. These means are finite: categorical outputs have finite ranges, and the Lipschitz bound for a continuous structural function bounds its output by its value at a fixed input plus L times the input distance; induction through the finite graph, using bounded non-Gumbel noises, gives finite first moments at every fixed (a, t) . Independence of the descendant noises from (A, T) shows that $g_{\mathcal{N}}$ is a version of the observational outcome regression given (A, T) . The intervention leaves the equations generating A unchanged, so its distribution is its observational marginal. Therefore,

$$\mathbb{E}_{\mathcal{M}}[Y(t)] = \int g_{\mathcal{M}}(a, t) \nu_A(da), \quad \mathbb{E}_{\widehat{\mathcal{M}}}[\widehat{Y}(t)] = \int g_{\widehat{\mathcal{M}}}(a, t) \mu_A(da),$$

where ν_A and μ_A are the observational marginals of A .

To bound the Lipschitz constants, take an arbitrary $\mathcal{N} \in \{\mathcal{M}, \widehat{\mathcal{M}}\}$. For a continuous descendant V_i , couple its outputs at two parent histories h, h' using the same private noise. The structural Lipschitz bound gives

$$\mathbb{E}\|V_i(h) - V_i(h')\| \leq L\|h - h'\|.$$

For a categorical descendant, let $\pi_i(h)$ be its propensity vector after integrating its private non-Gumbel noise. The canonical Gumbel rule has categorical probabilities equal to this vector. Integration of the propensity Lipschitz bound gives

$$\|\pi_i(h) - \pi_i(h')\| \leq L\|h - h'\|.$$

Place mass $\min\{\pi_{i,j}(h), \pi_{i,j}(h')\}$ on each equal pair (j, j) . The remaining masses in the two marginals have the same total mass and disjoint supports; coupling these remaining masses gives disagreement probability

$$1 - \sum_{j=1}^{n_i} \min\{\pi_{i,j}(h), \pi_{i,j}(h')\} = \frac{1}{2} \sum_{j=1}^{n_i} |\pi_{i,j}(h) - \pi_{i,j}(h')|.$$

Two labels in $[n_i]$ differ by at most $n_i - 1$, so this coupling satisfies

$$\mathbb{E}|V_i(h) - V_i(h')| \leq \frac{n_i - 1}{2} \|\pi_i(h) - \pi_i(h')\| \leq LD\|h - h'\|.$$

The continuous bound is also at most $LD\|h - h'\|$, since $D \geq 1$.

Starting at fixed inputs (a, t) and (a', t') , use these couplings successively for the descendants, conditional on the accumulated histories. Let E_r be the expected 1-norm distance between these

histories after r descendants have been generated. Projection onto a parent history cannot increase this distance. Hence

$$E_0 = \|a - a'\| + |t - t'|, \quad E_{r+1} \leq E_r + LDE_r = (1 + LD)E_r.$$

Induction gives $E_r \leq (1 + LD)^r E_0$: the assertion holds at $r = 0$, and the recurrence multiplies its bound by $1 + LD$ at the next step. There are at most n_V descendants. Projection onto the outcome coordinate, followed by $|\mathbb{E}Y - \mathbb{E}Y'| \leq \mathbb{E}|Y - Y'|$, therefore gives

$$|g_{\mathcal{N}}(a, t) - g_{\mathcal{N}}(a', t')| \leq L_y(\|a - a'\| + |t - t'|).$$

This argument applies to each of the two models without conditioning on latent inputs shared with observed ancestors.

Let Q be the set of allowed inputs $s = (a, t)$ whose numerical coordinates all lie in $[-K, K]$. Categorical coordinates retain their categorical ranges. The numerical dimension of (A, T) is at most q , so any $s, s' \in Q$ satisfy $\|s - s'\| \leq 2Kq$. Both observational laws of (A, T) give Q probability one. For each $\mathcal{N} \in \{\mathcal{M}, \widehat{\mathcal{M}}\}$, the observational outcome bound and the conditional expectation identity give $|g_{\mathcal{N}}(A, T)| \leq K$ almost surely under its own observational law. Choose an input $s_{\mathcal{N}} \in Q$ where this inequality holds. For every $s \in Q$, the Lipschitz inequality then yields

$$|g_{\mathcal{N}}(s)| \leq |g_{\mathcal{N}}(s_{\mathcal{N}})| + L_y\|s - s_{\mathcal{N}}\| \leq K + 2KqL_y = B.$$

Thus the structural regression versions are bounded by B on Q , including inputs outside their observational supports. Put $f_y(a, t) = g_{\mathcal{M}}(a, t)$. Recall that ν, μ are the observation distributions of $\mathcal{M}$ and $\widehat{\mathcal{M}}$ respectively in Proposition 4.3. Put $\widehat{f}_y(a, t) = g_{\widehat{\mathcal{M}}}(a, t)$. Its construction gives a version of $\mathbb{E}_{\widehat{\mathcal{M}}}[\widehat{Y}|A = a, T = t]$, the L_y -Lipschitz inequality, the interventional identity, and the bound $|\widehat{f}_y(a, t)| \leq B$ for $(a, t) \in Q$. We use the test-function argument of Lemma C.3 on the vector covariate $s = (a, t)$, with regression bound B , and give the calculation explicitly. Write

$$d(s) = \widehat{f}_y(s) - f_y(s), \quad G(s, y) = d(s)y, \quad h(s) = d(s)\widehat{f}_y(s).$$

On Q , the established bounds imply $|d(s)| \leq 2B$ and $|d(s) - d(s')| \leq 2L_y\|s - s'\|$. For

$s, s' \in Q$ and $|y|, |y'| \leq K \leq B$,

$$\begin{aligned}
|G(s, y) - G(s', y')| &\leq 2B|y - y'| + 2KL_y\|s - s'\| \\
&\leq 2B \max\{L_y, 1\}(\|s - s'\| + |y - y'|), \\
|h(s) - h(s')| &\leq |d(s)| |\widehat{f}_y(s) - \widehat{f}_y(s')| + |\widehat{f}_y(s')| |d(s) - d(s')| \\
&\leq 4BL_y\|s - s'\|.
\end{aligned}$$

Conditioning on the covariate gives $\mathbb{E}_\mu[G] = \mathbb{E}_\mu[h]$ and $\mathbb{E}_\nu[G] = \mathbb{E}_\nu[df_y]$. Since $h - df_y = d^2$,

$$\begin{aligned}
\mathbb{E}_\nu[d(A, T)^2] &= (\mathbb{E}_\mu[G] - \mathbb{E}_\nu[G]) + (\mathbb{E}_\nu[h] - \mathbb{E}_\mu[h]) \\
&\leq (2B \max\{L_y, 1\} + 4BL_y)W(\mu, \nu) = C_W W(\mu, \nu).
\end{aligned}$$

For each Wasserstein bound, integrate the corresponding displayed Lipschitz inequality against any coupling of the observational laws and take the infimum of its expected 1-norm cost. Both endpoints belong to the domains on which these inequalities were proved. Projection onto (A, T, Y) or (A, T) cannot increase that cost.

Set $I = [t_1, t_2]$. Observational boundedness gives $\nu_T(\mathbb{R} \setminus [-K, K]) = 0$. The conditional treatment law given A equals the law given $X = \text{Pa}(T)$. Integrating the overlap bound therefore gives

$$0 = \nu_T(I \setminus [-K, K]) \geq \delta P(I \setminus [-K, K]),$$

so $P(I \setminus [-K, K]) = 0$. Thus, for P -almost every $t \in I$ and for either observational marginal of A , the input (A, t) belongs to Q . The overlap assumption now yields

$$\begin{aligned}
C_W W(\mu, \nu) &\geq \int (f_y(a, t) - \widehat{f}_y(a, t))^2 \nu(dt|A = a) \nu_A(da) \\
&\geq \delta \int_I \int (f_y(a, t) - \widehat{f}_y(a, t))^2 \nu_A(da) P(dt) \\
&\geq \delta \int_I b(t)^2 P(dt),
\end{aligned}$$

where $b(t) = \int (f_y(a, t) - \widehat{f}_y(a, t)) \nu_A(da)$. The last inequality is Jensen's inequality for the probability measure ν_A . Put

$$c(t) = \int \widehat{f}_y(a, t) \nu_A(da) - \int \widehat{f}_y(a, t) \mu_A(da).$$

For P -almost every $t \in I$, integrate the Lipschitz bound of $\widehat{f}_y$ in a against a coupling of the two marginals and take the infimum to obtain $|c(t)| \leq L_y W(\nu_A, \mu_A) \leq L_y W(\mu, \nu)$. The interventional identities give

$$\mathbb{E}_{\mathcal{M}}[Y(t)] - \mathbb{E}_{\widehat{\mathcal{M}}}[\widehat{Y}(t)] = b(t) + c(t).$$

Since $2b(t)c(t) \leq b(t)^2 + c(t)^2$, it follows, for P -almost every $t \in I$, that

$$(\mathbb{E}_{\mathcal{M}}[Y(t)] - \mathbb{E}_{\widehat{\mathcal{M}}}[\widehat{Y}(t)])^2 \leq 2b(t)^2 + 2L_y^2 W^2(\mu, \nu).$$

Integrating this inequality against $P(dt)$ on $[t_1, t_2]$ and using the bound on the integral of $b(t)^2$ gives

$$\begin{aligned} \int_{t_1}^{t_2} (\mathbb{E}_{\mathcal{M}}[Y(t)] - \mathbb{E}_{\widehat{\mathcal{M}}}[\widehat{Y}(t)])^2 P(dt) &\leq 2 \int_{t_1}^{t_2} b(t)^2 P(dt) + 2L_y^2 W^2(\mu, \nu) \int_{t_1}^{t_2} 1 P(dt) \\ &\leq \frac{2C_W}{\delta} W(\mu, \nu) + 2L_y^2 W^2(\mu, \nu) P([t_1, t_2]). \end{aligned}$$

□

Proof of Corollary 4.4. We establish feasibility and the complete-observation sampling bounds for the specified restricted optimization. Write η for a network representation, θ_η for its population SCM parameter, and θ^* for the true parameter. For two population parameters define

$$d(\theta, \theta') = \sum_{i=1}^{n_V} \|f_i - f'_i\|_\infty + \sum_{j=1}^{n_U} W(P(U_j), P(U'_j)),$$

using normalized categorical propensity functions in the first sum. Let $\mathcal{A}_n$ be the set of admissible representations. With common independent exogenous draws $\xi_1, \dots, \xi_{m_n}$, let $H_\eta(\xi_l)$ be the complete observed vector and put

$$P_{m_n}^\eta = \frac{1}{m_n} \sum_{l=1}^{m_n} \delta_{H_\eta(\xi_l)}, \quad h_n(\eta) = S_{\lambda_n}(P_n^{\theta^*}, P_{m_n}^\eta), \quad g(\theta) = W(P^{\theta^*}, P^\theta).$$

All distributions in these expressions are distributions of the complete observed vector. Write $f(\theta) = \mathbb{E}_{t \sim \mu_T} \mathbb{E}_{\mathcal{M}^\theta}[F(\mathbf{V}(t))]$ for the objective.

For fixed n , pad the finitely many network architectures to their stated widths and encode each representation by its finitely many real weights and mixture probabilities. Finite products of

Euclidean spaces and simplices are standard Borel spaces. Affine maps, ReLU, clipping and finite composition have jointly continuous evaluation in their weights and inputs. At the fixed temperature $\tau_n > 0$, mixture weights can be written

$$\frac{p_j^{1/\tau_n} \exp(G_j/\tau_n)}{\sum_k p_k^{1/\tau_n} \exp(G_k/\tau_n)}, \quad p \in \Delta,$$

whose denominator is positive even when some $p_j = 0$; hence generator evaluation is jointly Borel.

The restrictions defining $\mathcal{A}_n$ are Borel restrictions on these parameters. Choose a countable dense set in each continuous structural input domain. Continuity makes the output bound, coordinatewise nonnegativity and every Lipschitz inequality equivalent to their inequalities on that dense set or its Cartesian square, and each such inequality is Borel in the weights. For a nonnegative categorical raw network q , positivity on a compact input box D_0 is equivalent to membership in

$$\bigcup_{l=1}^{\infty} \bigcap_{x \in Q_0} \{\eta : \|q_\eta(x)\|_1 \geq 1/l\},$$

where Q_0 is countable and dense in D_0 . Positivity implies a positive minimum on the compact box; conversely the dense-set lower bound extends by continuity. Intersect these conditions over a countable compact exhaustion if the full domain is unbounded. On the resulting Borel parameter set $q/\|q\|_1$ is jointly Borel and continuous in the input, so its Lipschitz restriction is another countable family of Borel inequalities. The raw Lipschitz restrictions before clipping have the same dense-input characterization. Thus $\mathcal{A}_n$ is standard Borel.

The exogenous space is a finite product of uniform cubes and real Gumbel coordinates. For categorical propensities p , define $r_i = \exp(-g_i)/p_i$ for $p_i > 0$ and $r_i = \infty$ for $p_i = 0$. The smallest minimizing label is i precisely when $r_i < r_j$ for all $j < i$ and $r_i \leq r_j$ for all $j > i$. These are Borel comparisons of extended-real Borel functions. Induction in topological order proves that both $H_\eta(\xi)$ and $H_\eta(t, \xi)$ are jointly Borel. Integration of the bounded objective on the observed range makes $f(\theta_\eta)$ Borel in η : parameter integrals of bounded Borel functions are Borel by the monotone class theorem. Empirical transport values are continuous in the cost matrix, since a uniform change of at most a in costs changes every coupling objective and its infimum by at most a . Hence h_n is jointly Borel in parameters and data. The event $\{F_n < a\}$ is the projection of the Borel set $\{(\eta, \text{data}) : h_n(\eta) \leq \alpha_n, f(\theta_\eta) < a\}$, so is analytic and measurable in the completed sample law.

First derive population continuity in the fixed dimensions. Couple corresponding independent non-Gumbel latent laws optimally and take the product of their couplings. The expected joint latent 1-norm discrepancy is the sum of their marginal Wasserstein distances. Generate the observed nodes in topological order. At a continuous node, the expected new discrepancy is at most $d_i \|f_i - f'_i\|_\infty + \hat{L}_f$ times the expected input discrepancy. At a categorical node, place mass $\min\{p_k, p'_k\}$ on each equal label pair and couple the remaining disjoint residual masses. Its disagreement probability is $\|p - p'\|_1/2$, so its expected label discrepancy is at most $(n_i - 1)\|p - p'\|_1/2$. The normalized Lipschitz restrictions bound this by

$$\frac{n_i(n_i - 1)}{2} \|f_i - f'_i\|_\infty + \frac{(n_i - 1)\hat{L}_f}{2} \mathbb{E} \|\text{input} - \text{input}'\|.$$

Choose fixed $A \geq 1$ and B bounding these finitely many coefficients. If e_k is the expected discrepancy of the latent vector and generated history, then $e_0 \leq d(\theta, \theta')$ and $e_{k+1} \leq Ae_k + Bd(\theta, \theta')$. Induction gives

$$e_k \leq A^k e_0 + Bd(\theta, \theta') \sum_{r=0}^{k-1} A^r :$$

the assertion is equality at $k = 0$, and multiplying its right-hand side by A and adding $Bd(\theta, \theta')$ gives the asserted expression at $k + 1$. Projection onto the observed vector yields $W(P^\theta, P^{\theta'}) \leq Cd(\theta, \theta')$. Identical intervention values contribute zero discrepancy and remove their generation steps, giving the same bound for every intervention. The triangle inequality consequently makes g Lipschitz; denote a Lipschitz constant by C_g .

We now bound the complete observed simulation error, retaining paired observations and all categorical noise. Put $H_n = 1 + L_{1,n} + L_{2,n}$. The component counts of the specified separable architectures are fixed. For $0 < \rho < 1/2$, round the joint non-Gumbel latent vector to a grid of 1-norm mesh at most ρ in its D_U -dimensional cube. There are at most $C\rho^{-D_U}$ grid points. For each grid point and each possible categorical prefix in a topological order, evaluate the structural equations with that prefix fixed. Record the continuous outputs and categorical propensity vectors in a table, and round them within ρ , using a finite simplex grid for propensities. There are finitely many prefixes, independent of n . The tables have a finite set $\mathcal{T}_\rho$ of possible values with

$$\log |\mathcal{T}_\rho| \leq C\rho^{-D_U} \log(2 + \rho^{-1}).$$

Indeed, each entry lies in a fixed bounded finite-dimensional box, so a grid has at most $(C/\rho)^C$ choices, and there are at most $C\rho^{-D_U}$ entries. With a categorical prefix fixed, induction through the continuous structural equations bounds their change under a latent perturbation of size ρ by $C\rho$: each new difference is at most $\hat{L}_f$ times the sum of previous and latent differences. Categorical propensities satisfy the same bound. Thus, while the selected categorical prefixes agree, the original and table outputs differ by at most $C\rho$, and corresponding propensities differ coordinatewise by at most $a = C\rho$.

The table simulation uses the original categorical Gumbel draws. Write $E_k = \exp(-g_k)$; these are independent unit-rate exponential variables, and a propensity vector p selects $\arg \min_k E_k/p_k$, with $E_k/0 = \infty$. For a table propensity vector $q \in \Delta_N$, call label i robust if $q_i > a$ and

$$\frac{E_i}{q_i - a} < \min_{j \neq i} \frac{E_j}{q_j + a}.$$

A robust label wins for every propensity vector whose coordinates differ from q by at most a . The robust-label events are disjoint. When $q_i > a$, integration of the independent exponential densities gives

$$\mathbb{P}(i \text{ robust}) = \frac{q_i - a}{q_i - a + \sum_{j \neq i} (q_j + a)}.$$

When $q_i \leq a$, the probability is zero. In either case it is at least $(q_i - a)_+ / (1 + Na)$. Summing and using $\sum_i (q_i - a)_+ \geq 1 - Na$ bounds the probability of no robust label by $2Na / (1 + Na)$ when $Na < 1$; when $Na \geq 1$, the bound $1 \leq 2Na$ suffices. For each table and generator representation let b indicate that a nonrobust choice occurs at any table prefix. Independence of categorical noise from the generated latent vector and a union bound over the finitely many prefixes give $\mathbb{E}b \leq C\rho$. If $b = 0$, induction over the categorical nodes forces the original and table prefixes to agree. Thus, with $\tilde{H}$ the table observed map and B_V a common observed 1-norm diameter bound,

$$\|H_\eta(\xi) - \tilde{H}(\xi)\| \leq C\rho + B_V b(\xi).$$

This bound includes zero propensities.

A table simulation has type (u, c) , where u is its latent grid index and c its complete categorical label vector; its observed output is fixed for each type. There are at most $J = C\rho^{-D_U}$ types. Include the indicators of membership in every subset of these types, and the ambiguity indicators b , in a scalar class $\mathcal{B}_{n,\rho}$.

To verify the structural condition of Lemma E.6 for this class, let Γ_n be the entire finite-dimensional generator weight space with its mixture simplex factors. This gives a measurable envelope for the restricted generator representations. Index indicators by a parameter in Γ_n , a table in the finite set $\mathcal{T}_\rho$, and a finite tag specifying either a subset of the at most J types or the ambiguity indicator. This is a standard Borel space. Generator evaluation is jointly Borel by the positive-denominator mixture formula. Its clipped output has a Borel grid index when cells are half-open and the final cell includes its endpoint. Given the table and grid index, each categorical choice is a finite Borel comparison using $r_i < r_j$ for $j < i$ and $r_i \leq r_j$ for $j > i$, with $r_i = \infty$ for zero table propensity. Each robust event is a Borel comparison with positive denominators on $q_i > a$, and is empty on $q_i \leq a$. Finite unions and complements therefore give jointly Borel ambiguity and type-subset indicators. This includes all ties, grid boundaries and zero propensities. Clip uniform exogenous coordinates outside their cubes to view these functions on the ambient Euclidean input space; the values on the sampled inputs do not change. Composing a jointly Borel indicator with a continuous scalar dual function and its normalized square-root transform preserves joint Borel evaluation. Thus all the classes to which the lemma's concentration argument is applied have the required standard Borel parameterization. The envelope uses the same architectural parameters and finite tags as the growth calculation; no density assertion for discontinuous selectors is needed.

On M_0 fixed exogenous inputs, the latent grid indices are determined by at most CM_0/ρ scalar generator comparisons. The ReLU blocks have at most CH_n parameters and CH_n layers. Within a fixed activation pattern their outputs are polynomials in the parameters of degree at most CH_n . For a positive-temperature mixture write its weights as $a_j R_j / \sum_k a_k R_k$, where $a_j = p_j^{1/\tau_n} \geq 0$, $\sum_j a_j > 0$, and $R_j = \exp(G_j/\tau_n)$ is fixed on each input. Treat the a_j as additional free parameters. Clearing the positive denominator in a grid comparison leaves a polynomial of degree at most $CH_n + 1$. Truncation adds fixed-degree comparisons. The polynomial sign bound in Lemma 17 of [6], applied successively to activation comparisons in each layer and then to grid comparisons, gives at most

$$(C(M_0 + H_n)H_n/\rho)^{CH_n^2}$$

latent grid patterns: each layer contributes at most $(C(M_0 + H_n)H_n/\rho)^{CH_n}$, and there are at most CH_n layers. For a fixed table and grid pattern all categorical race comparisons and table outputs on these inputs are determined. The growth of $\mathcal{B}_{n,\rho}$ is therefore at most this number times

$|\mathcal{T}_\rho|2^J$. Hence its pseudo-dimension satisfies

$$v_n(\rho) \leq C \left(\rho^{-D_U} \log(2 + \rho^{-1}) + H_n^2 \log(2 + H_n/\rho) \right).$$

To obtain this bound explicitly, put $A_0 = \log |\mathcal{T}_\rho| + J \log 2$ and $B_0 = CH_n^2$ in the shattering inequality

$$M_0 \log 2 \leq A_0 + B_0 \log(C(M_0 + H_n)H_n/\rho).$$

If $M_0 \leq 4A_0/\log 2$, the first term suffices. Otherwise the inequality bounds M_0 by $CB_0 \log(2 + B_0 H_n/\rho) + CB_0 \log(2 + A_0/B_0)$; the established bound on A_0 gives the displayed estimate.

Put

$$r_n = m_n^{-1/(D+2)} \log m_n + H_n m_n^{-1/2} \log m_n.$$

The depth schedule in Theorem 4.2 gives $H_n = O(m_n^{d_{\max}^U/(2d_{\max}^U+4)})$, so $r_n \rightarrow 0$. For sufficiently large n take $\rho = r_n$ and apply Lemma E.6 to $\mathcal{B}_{n,r_n}$ with scalar output. These functions take values in $\{0, 1\}$, and the Wasserstein distance of their Bernoulli output laws equals the absolute difference of their means. With tolerances $\delta = r_n$ and $\epsilon = Cr_n$, its failure exponent is bounded above by

$$C_1 \left((1 + r_n^{-1}) \log(2 + r_n^{-1}) + r_n^{-D_U} \log^2(2 + r_n^{-1}) + H_n^2 \log(2 + H_n/r_n) \log(2 + r_n^{-1}) \right) - cC^2 m_n r_n^2.$$

Here $\log(2 + H_n/r_n) = O(\log m_n)$ and

$$m_n r_n^2 \geq m_n^{D/(D+2)} \log^2 m_n + H_n^2 \log^2 m_n,$$

whereas $r_n^{-D_U} \leq m_n^{D_U/(D+2)} (\log m_n)^{-D_U}$ and $r_n^{-1} \leq m_n^{1/(D+2)} / \log m_n$. Since $D \geq \max\{D_U, 1\}$, a sufficiently large fixed C makes the failure probability $O(m_n^{-3})$. On the complementary event every type-subset probability and ambiguity-indicator mean differs from its empirical counterpart by at most Cr_n . The type-subset bound controls total variation between population and empirical type laws. Couple common type masses identically and residual masses arbitrarily; their table-output Wasserstein distance is at most $B_V Cr_n$. Coupling original and table outputs on the same draw bounds their population discrepancy by $C\rho + B_V \mathbb{E}b$ and their empirical discrepancy by $C\rho + B_V m_n^{-1} \sum_l b(\xi_l)$. Both are at most Cr_n by the ambiguity estimate. The triangle inequality therefore gives

$$\sup_{\eta \in \mathcal{A}_n} W(P^{\theta_\eta}, P_{m_n}^\eta) \leq Cr_n$$

with probability at least $1 - O(m_n^{-3})$.

For the observational sample let C be its complete categorical label vector, taking values in a fixed finite set. Conditional on $C = c$, all continuous observations equal a deterministic map $\Phi_c(U_1, \dots, U_{n_U})$, obtained by fixing the categorical labels and evaluating the continuous equations in topological order. Their Lipschitz bounds make Φ_c Lipschitz with a constant uniform over these finitely many labels. The conditional joint latent laws $\mu_c = P((U_1, \dots, U_{n_U}) \mid C = c)$ are supported on the bounded D_U -dimensional cube and need not be products. Write $p_c = P(C = c)$ and N_c for sample counts. If $p_c = 0$, then $N_c = 0$ almost surely. For $p_c > 0$, binomial concentration gives $N_c \geq np_c/2$ except on an exponentially small event. Conditional on the sample labels, latents in category c are independent with law μ_c .

For $D_U \geq 1$, Theorem 2 of [17] applied to these conditional joint latent laws bounds their empirical Wasserstein errors by $C_c n^{-1/\max\{2, D_U\}} \log^2 n$ with failure probability $O(n^{-4})$, uniformly for $np_c/2 \leq N_c \leq n$. Indeed, its small-tolerance exponents are constant multiples of $N_c x^2$ for $D_U = 1$, $N_c(x/\log(2+1/x))^2$ for $D_U = 2$, and $N_c x^{D_U}$ for $D_U > 2$. At the specified tolerance these grow at least as $\log^4 n$, $\log^2 n$, and $\log^{2D_U} n$, respectively. Bounded support supplies the exponential-moment assumption. A coupling pushed through Φ_c multiplies its cost by at most the Lipschitz constant of Φ_c . For $D_U = 0$, each conditional observed law is a point mass and its conditional empirical error is zero. Binomial concentration also gives $\sum_c |N_c/n - p_c| \leq C\sqrt{\log n/n}$ with probability at least $1 - O(n^{-4})$. Couple common categorical-component masses by their conditional couplings and unmatched masses arbitrarily within the observed support. After a union bound over the finitely many categories,

$$W(P_n^{\theta^*}, P^{\theta^*}) \leq \sum_c \min\{p_c, N_c/n\} W((\Phi_c)_\# \mu_c, (\Phi_c)_\# (\mu_c)_{N_c}) + \frac{B_V}{2} \sum_c |N_c/n - p_c| \leq t_n,$$

$$t_n = C_2 n^{-1/\max\{2, D_U\}} \log^2 n,$$

with probability at least $1 - O(n^{-3})$. Terms with zero common mass are omitted. Lemma E.5 and the triangle inequality give, on the combined event,

$$\sup_{\eta \in \mathcal{A}_n} |h_n(\eta) - g(\theta_\eta)| \leq 2(1 + \log(nm_n))\lambda_n + Cr_n + t_n \leq s_n.$$

The event has probability at least $1 - O(n^{-3} + m_n^{-3})$.

We construct a feasible approximation to the true model. For a structural function $f : [-K, K]^m \rightarrow \mathbb{R}^{m'}$ with $m \geq 1$, divide every coordinate into R intervals of length $h = 2K/R$ and triangulate each cell into simplices with successive vertices $v_0, v_1 = v_0 + he_{\pi(1)}, \dots, v_m = v_{m-1} + he_{\pi(m)}$. Let $I_h f$ interpolate the vertex values affinely. Its derivative column in direction $e_{\pi(j)}$ is $(f(v_j) - f(v_{j-1}))/h$, of 1-norm at most L_f . Subdividing a segment at simplex boundaries and summing these derivative bounds proves $\text{Lip}(I_h f) \leq L_f$. With barycentric coordinates a_j ,

$$\|I_h f(x) - f(x)\| \leq \sum_{j=0}^m a_j \|f(v_j) - f(x)\| \leq L_f m h.$$

Convex interpolation preserves the bound K for continuous outputs and preserves the simplex for categorical propensities. Its nodal basis functions are translates and rescalings of

$$\phi(z) = \max \{0, \min (\{1 + z_k, 1 - z_k : k \in [m]\} \cup \{1 + z_k - z_l : k, l \in [m]\})\}.$$

On each simplex of a nodal star this function is affine, equals one at the central vertex, and equals zero at its other vertices. Hence $I_h f(x) = \sum_v f(v) \phi((x - v)/h)$. Compute each minimum by a binary tree using $\min(a, b) = a - \max\{a - b, 0\}$. Each term has width $O(m^2)$ and depth $O(1 + \log m)$. Parallelizing $O(R^m)$ terms and summing their outputs gives the required fixed-dimensional depth and width $O(R^m)$. Coordinatewise input clipping is a contraction and preserves the Lipschitz bound. Taking R proportional to $W_{0,n}^{1/m}$ gives error $O(W_{0,n}^{-1/m})$. A function with $m = 0$ is constant and represented exactly. For categorical functions take $q = I_h f$; it is nonnegative and normalized, and its normalized Lipschitz constant equals its original one. Thus these approximants satisfy the categorical restrictions. Continuous output clipping preserves the error bound and cannot increase the Lipschitz constant.

For each normalized latent support component measure μ_i , let $\psi_i : [0, 1]^{q_i} \rightarrow C_i$ be its Lipschitz chart. For $q_i > 0$, Proposition B.2 applied to its pullback measure gives a curve $\gamma_i : [0, 1] \rightarrow [0, 1]^{q_i}$ of Holder exponent $1/q_i$ with that pullback law. Theorem 2(b) of [52] applied to its coordinates gives first-block uniform error $O(L_{1,n}^{-2/q_i})$ and, applied to the Lipschitz coordinates of ψ_i , gives second-block error $O(L_{2,n}^{-2/q_i})$. Parallelize coordinate networks and clip first-block

outputs to their input cubes. Their widths fit the specified budgets. Lemma E.1 bounds the resulting component Wasserstein error by $O(L_{1,n}^{-2/q_i} + L_{2,n}^{-2/q_i})$. For $q_i = 0$, use the exact constant component. Lemma E.2 bounds the hard-selected mixture error by the weighted component errors. Couple positive-temperature and hard Gumbel weights, retaining identical independent bounded component outputs v_i . Then

$$\left\| \sum_i (X_i^{\tau_n} - X_i^0) v_i \right\| \leq C \|X^{\tau_n} - X^0\|_1.$$

Lemma E.4 bounds the additional error by $O(\tau_n(1 - \log \tau_n))$ for sufficiently large n . Output truncation cannot increase the latent error because the true latents lie in $[-K, K]^{d_j^U}$ and truncation contracts distance to every point of this cube. Choose the latent generators independently. Summing their marginal errors and the structural errors gives a representation $\eta_n \in \mathcal{A}_n$ satisfying

$$a_n = d(\theta^*, \theta_{\eta_n}) \leq O \left(W_{0,n}^{-1/d_{\max}^{\text{in}}} + \sum_{i=1}^2 L_{i,n}^{-2/d_{\max}^U} + \tau_n(1 - \log \tau_n) \right).$$

Choose the constant in ϵ_n so that $C_g a_n \leq \epsilon_n$. On the combined sampling event,

$$h_n(\eta_n) \leq g(\theta_{\eta_n}) + s_n \leq C_g a_n + s_n \leq \epsilon_n + s_n = \alpha_n.$$

Thus the optimization is feasible. Since $m_n = \Omega(n)$, the same event has probability at least $1 - O(n^{-2})$ and gives

$$W(P_n^{\mathcal{M}^*}, P^{\mathcal{M}^*}) \leq O(\alpha_n), \quad \sup_{\eta \in \mathcal{A}_n} W(P_{m_n}^\eta, P^{\theta_\eta}) \leq O(\alpha_n).$$

For every feasible representation η , Lemma E.5 and the empirical constraint give

$$W(P_n^{\mathcal{M}^*}(\mathbf{V}), P_{m_n}^\eta(\mathbf{V})) \leq S_{\lambda_n}(P_n^{\mathcal{M}^*}(\mathbf{V}), P_{m_n}^\eta(\mathbf{V})) + 2(1 + \log(nm_n))\lambda_n \leq O(\alpha_n),$$

since the regularization term is included in $s_n \leq \alpha_n$. On the same sampling event, the triangle inequality gives, for every feasible representation η ,

$$\begin{aligned} w_\eta &= W(P_n^{\mathcal{M}^*}(\mathbf{V}), P^{\theta_\eta}(\mathbf{V})) \\ &\leq W(P_n^{\mathcal{M}^*}(\mathbf{V}), P_n^{\mathcal{M}^*}(\mathbf{V})) + W(P_n^{\mathcal{M}^*}(\mathbf{V}), P_{m_n}^\eta(\mathbf{V})) + W(P_{m_n}^\eta(\mathbf{V}), P^{\theta_\eta}(\mathbf{V})) \\ &\leq O(\alpha_n). \end{aligned}$$

All feasible models have bounded continuous structural functions and normalized categorical propensity functions with Lipschitz constant at most $\widehat{L}_f$. The true model also belongs to the corresponding bounded Lipschitz class. Apply Proposition 4.3 to the true model and each feasible model, using their common unconfounded graph and the overlap assumption of the true model. Write

$$b_\eta(t) = \mathbb{E}_{\mathcal{M}^*}[Y(t)] - \mathbb{E}_{\mathcal{M}^{\theta_\eta}}[Y(t)], \quad r(t) = \frac{d\mu_T}{dP}(t).$$

With the fixed constants C_W, L_y supplied by Proposition 4.3 at Lipschitz bound $\widehat{L}_f$, it gives

$$\int_{t_1}^{t_2} b_\eta(t)^2 P(dt) \leq \frac{2C_W}{\delta} w_\eta + 2L_y^2 w_\eta^2 P([t_1, t_2]).$$

Absolute continuity and Cauchy–Schwarz therefore give

$$\begin{aligned} |\mathbb{E}_{t \sim \mu_T} \mathbb{E}_{\mathcal{M}^{\theta_\eta}}[Y(t)] - F_*| &= \left| \int_{t_1}^{t_2} b_\eta(t) r(t) P(dt) \right| \\ &\leq \left(\int_{t_1}^{t_2} b_\eta(t)^2 P(dt) \right)^{1/2} \left(\int_{t_1}^{t_2} r(t)^2 P(dt) \right)^{1/2} \\ &\leq A_T \left(\frac{2C_W}{\delta} w_\eta + 2L_y^2 w_\eta^2 P([t_1, t_2]) \right)^{1/2} \\ &\leq C \sqrt{\alpha_n}. \end{aligned}$$

For the last inequality, the growth assumptions imply $\alpha_n \rightarrow 0$, so $\alpha_n \leq 1$ eventually and $w_\eta^2 = O(\alpha_n^2) = O(\alpha_n)$; all other factors are fixed. The bound is uniform over feasible representations. Feasibility was established on the same event, so every feasible objective lies in $[F_* - C\sqrt{\alpha_n}, F_* + C\sqrt{\alpha_n}]$. Taking its infimum proves $|F_n - F_*| \leq C\sqrt{\alpha_n}$, whether or not the infimum is attained. The event has probability at least $1 - O(n^{-2})$. $\square$

D Experiments

The structure equations of the generative models are (3.2). We use three-layer feed-forward neural networks with width 128 for each $\widehat{f}_i$ and six-layer neural networks with width 128 for each $\widehat{g}_j$. We use the Augmented Lagrangian Multiplier (ALM) method to solve the optimization problems as in [3]. We run 600 epochs and use a batch size of 2048 in each epoch. m_n is set to be $m_n = n$. The "geomloss" package [16] is used to calculate the Sinkhorn distance. To impose Lipschitz

regularization, we use the technique from [23] to do layer-wise normalization to the weight matrices with respect to infinity norm. The upper bound of the Lipschitz constant in each layer is set to be 8. The τ in the Gumbel-softmax layer is set to be 0.

For the choice of Wasserstein ball radius α_n , we use the subsampling technique from [11, Section 3.4]. We can take the criterion function in [11] to be $Q(\theta) = W(P^{\mathcal{M}^*}(\mathbf{V}), P^{\mathcal{M}^\theta}(\mathbf{V}))$ and its empirical estimation to be $Q_n(\theta) = W(P_n^{\mathcal{M}^*}(\mathbf{V}), P_n^{\mathcal{M}^\theta}(\mathbf{V}))$. In the first step, we minimize the Wasserstein distance $\hat{\theta}_n^* = \arg \min_{\theta} W(P_n^{\mathcal{M}^*}(\mathbf{V}), P_n^{\mathcal{M}^\theta}(\mathbf{V}))$. Then, [11] propose to refine the radius $Q_n(\hat{\theta}_n^*)$ by taking the quantile of subsample $\{\sup_{\theta: Q_n(\hat{\theta}) \leq \beta Q_n(\hat{\theta}_n^*)} Q_{j,b}(\theta)\}$, where $Q_{j,b}$ denotes the criterion function estimated at j -th subsample of size b . However, it is time-consuming to solve this optimization problem many times. In practice, we set the radius α_n to be the 95% quantile of $\{W(\bar{P}_j^{\mathcal{M}^*}(\mathbf{V}), P_n^{\mathcal{M}^{\hat{\theta}_n^*}}(\mathbf{V}))\}$, where we use 50 subsamples $\bar{P}_j^{\mathcal{M}^*}, j = 1, \dots, 50$ from $P^{\mathcal{M}^*}(\mathbf{V})$ with size $b = \lfloor 15\sqrt{n} \rfloor$.

D.1 Counterexample

We test our neural causal method on the counterexample in Section C.1. We choose the noise U_y to be the normal variable. The structure equations are

$$P(U_1 = 1) = 3/5, P(U_1 = 0) = 2/5, P(U_2 = 1) = 1/3, P(U_2 = 0) = 2/3,$$

$$X = \begin{cases} -1 & , \text{w.p. } 1/2, \\ 0 & , \text{w.p. } 1/2, \end{cases}$$

$$P(T = 1|X = x) = \begin{cases} 1/2 & , x = -1, \\ 5/8 & , x = 0, \end{cases}$$

$$Y = \begin{cases} T + U_y & , X = -1, \\ U_1 + U_y & , X = 0, T = 1, \\ U_2 + U_y & , X = 0, T = 0, \end{cases}$$

We compare the confidence intervals of the unregularized neural casual method and the Lipschitz regularized one under different architectures. We choose the layerwise Lipschitz constants upper bound to be 5 and 1.2. As a benchmark, we also include the bound obtained by the Double Machine Learning estimator using the EconML package. The result is shown in Section D.1. For this

identifiable case example, the double ML estimator produces better bounds for the ATE. The intervals given by regularized and unregularized NCM are similar to the regularized one slightly better in the left figure, where we use medium-sized NNs. However, in the right figure, the obtained intervals after regularization are tighter, although slightly biased, compared with the regularized setting. From these two experiments, we conclude that the architecture of NNs will also influence the results and adding regularization during the training process can prevent extreme confidence intervals or inconsistency.

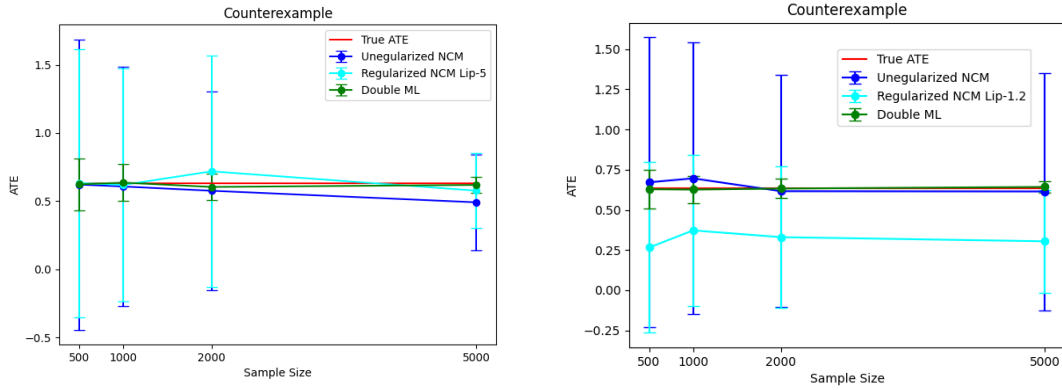


Figure 11: Comparison of Lipschitz regularized and unregularized neural causal algorithm. The two figures show the results of different architectures. The figure on the left side uses a medium-sized NN (width 128, depth 3) to approximate each structural function, while the right figure uses extremely small NNs (width 3, depth 1). In all experiments, we use the projected gradient to regularize the weight of the neural network. For each sample size, we repeat the experiment 5 times and take the average of the upper (lower) bound.

E Technical Lemmas

Lemma E.1. *Let $\mu, \hat{\mu}$ be two measures on compact set $K \subset \mathbb{R}^{d_1}$, for any measurable functions $F, \hat{F} : K \rightarrow \mathbb{R}^{d_2}$, if F is L -Lipschitz continuous, then*

$$W(F_{\#}\mu, \hat{F}_{\#}\hat{\mu}) \leq LW(\mu, \hat{\mu}) + d_2 \|F - \hat{F}\|_{\infty}.$$

Remark. The measures μ and $\hat{\mu}$ are probability measures. The uniform infinity norm is $\|F - \hat{F}\|_{\infty} = \sup_{x \in K} \max_{j \in [d_2]} |F_j(x) - \hat{F}_j(x)|$.

Proof. For any 1-Lipschitz function $g : \mathbb{R}^{d_2} \rightarrow \mathbb{R}$,

$$\begin{aligned} \mathbb{E}_{X \sim F_{\#}\mu}[g(X)] - \mathbb{E}_{X \sim \widehat{F}_{\#}\widehat{\mu}}[g(X)] &= \mathbb{E}_{X \sim \mu}[g \circ F(X)] - \mathbb{E}_{X \sim \widehat{\mu}}[g \circ \widehat{F}(X)] \\ &= (\mathbb{E}_{X \sim \mu}[g \circ F(X)] - \mathbb{E}_{X \sim \widehat{\mu}}[g \circ F(X)]) \\ &\quad + (\mathbb{E}_{X \sim \widehat{\mu}}[g \circ F(X)] - \mathbb{E}_{X \sim \widehat{\mu}}[g \circ \widehat{F}(X)]). \end{aligned}$$

Note that for any $x, y \in K$,

$$|g \circ F(x) - g \circ F(y)| \leq \|F(x) - F(y)\| \leq L\|x - y\|.$$

Thus $g \circ F$ is L -Lipschitz continuous and

$$\mathbb{E}_{X \sim \mu}[g \circ F(X)] - \mathbb{E}_{X \sim \widehat{\mu}}[g \circ F(X)] \leq LW(\mu, \widehat{\mu}). \quad (\text{E.1})$$

For the second term,

$$\begin{aligned} \mathbb{E}_{X \sim \widehat{\mu}}[g \circ F(X)] - \mathbb{E}_{X \sim \widehat{\mu}}[g \circ \widehat{F}(X)] &\leq \mathbb{E}_{X \sim \widehat{\mu}}|g(F(X)) - g(\widehat{F}(X))| \\ &\leq \mathbb{E}_{X \sim \widehat{\mu}}\|F(X) - \widehat{F}(X)\| \\ &\leq \sup_{x \in K} \|F(x) - \widehat{F}(x)\| \\ &\leq d_2\|F - \widehat{F}\|_{\infty}. \end{aligned} \quad (\text{E.2})$$

The last inequality follows by summing the d_2 coordinatewise bounds. If the uniform error is infinite, the asserted bound holds in the extended nonnegative reals. Combining (E.1) and (E.2) bounds the difference of expectations by $LW(\mu, \widehat{\mu}) + d_2\|F - \widehat{F}\|_{\infty}$ for every scalar 1-Lipschitz function g . Taking the supremum over these functions gives the asserted bound for $W(F_{\#}\mu, \widehat{F}_{\#}\widehat{\mu})$. $\square$

Lemma E.2. *Let $\mu_1, \dots, \mu_n, \widehat{\mu}_1, \dots, \widehat{\mu}_n$ be measure on $\mathbb{R}^d$, for any $p_1, \dots, p_n \in [0, 1]$ such that $\sum_{k=1}^n p_k = 1$, we have*

$$W\left(\sum_{k=1}^n p_k \mu_k, \sum_{k=1}^n p_k \widehat{\mu}_k\right) \leq \sum_{k=1}^n p_k W(\mu_k, \widehat{\mu}_k).$$

Additional condition. For every $k \in [n]$, $\int_{\mathbb{R}^d} \|x\| \mu_k(dx) + \int_{\mathbb{R}^d} \|x\| \widehat{\mu}_k(dx) < \infty$.

Remark. For every $k \in [n]$, μ_k and $\widehat{\mu}_k$ are probability measures on $\mathbb{R}^d$.

Remark (notation). μ_n denotes the locally specified measure, with n serving either as a component index or as the empirical sample size. In Lemma E.2: μ_n is the component measure with index n

in the mixture, among measures μ_k on $\mathbb{R}^d$ for $k \in [n]$. In Lemma E.3: μ_n is the final component measure in the product $\mu = \mu_1 \otimes \cdots \otimes \mu_n$. In Lemma E.6: $\mu_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ is the empirical distribution of the sample with law μ , on $\mathbb{R}^d$.

Proof. For any 1-Lipschitz function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we have

$$\begin{aligned} \mathbb{E}_{X \sim \sum_{k=1}^n p_k \mu_k} [f(X)] - \mathbb{E}_{X \sim \sum_{k=1}^n p_k \hat{\mu}_k} [f(X)] &= \sum_{k=1}^n p_k (\mathbb{E}_{X_k \sim \mu_k} [f(X_k)] - \mathbb{E}_{X_k \sim \hat{\mu}_k} [f(X_k)]) \\ &\leq \sum_{k=1}^n p_k W(\mu_k, \hat{\mu}_k). \end{aligned}$$

By dual formulation of Wasserstein-1 distance,

$$\begin{aligned} W\left(\sum_{k=1}^n p_k \mu_k, \sum_{k=1}^n p_k \hat{\mu}_k\right) &= \sup_{f \text{ 1-Lipschitz}} \mathbb{E}_{X \sim \sum_{k=1}^n p_k \mu_k} [f(X)] - \mathbb{E}_{X \sim \sum_{k=1}^n p_k \hat{\mu}_k} [f(X)] \\ &\leq \sum_{k=1}^n p_k W(\mu_k, \hat{\mu}_k). \end{aligned}$$

□

Lemma E.3. Given two measures $\mu = \mu_1 \otimes \cdots \otimes \mu_n$, $\hat{\mu} = \hat{\mu}_1 \otimes \cdots \otimes \hat{\mu}_n$, where every factor μ_k and $\hat{\mu}_k$ is a probability measure, we have

$$W(\mu, \hat{\mu}) \leq \sum_{k=1}^n W(\mu_k, \hat{\mu}_k).$$

Remark. W is Wasserstein-1 distance for the ground distance $\|x - y\|$. $S_\lambda(\mu, \nu) = W_{1,\lambda}(\mu, \nu) - \frac{1}{2}W_{1,\lambda}(\mu, \mu) - \frac{1}{2}W_{1,\lambda}(\nu, \nu)$.

Proof. If $n = 0$, both empty products are the probability measure on the one-point space and both sides are zero. Suppose $n \geq 1$. If any $W(\mu_k, \hat{\mu}_k) = \infty$, the inequality holds in the extended nonnegative reals. Otherwise, for any $\epsilon > 0$ and every $k \in [n]$, choose a probability coupling π_k of μ_k and $\hat{\mu}_k$ such that

$$\mathbb{E}_{(X_k, Y_k) \sim \pi_k} [\|X_k - Y_k\|_1] \leq W(\mu_k, \hat{\mu}_k) + \epsilon.$$

Such a coupling exists by the definition of the finite infimum defining W . Under the product probability law $\pi_1 \otimes \cdots \otimes \pi_n$, put $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_n)$. The marginal laws of X and Y are respectively μ and $\hat{\mu}$, because the coordinate pairs are independent and have the

prescribed marginal laws. Thus this construction is a coupling of the product measures. The product 1-norm is the sum of the coordinate 1-norms, so

$$\begin{aligned} W(\mu, \hat{\mu}) &\leq \mathbb{E}_{\pi_1 \otimes \dots \otimes \pi_n} [\|X - Y\|] \\ &= \sum_{k=1}^n \mathbb{E}_{(X_k, Y_k) \sim \pi_k} [\|X_k - Y_k\|_1] \\ &\leq \sum_{k=1}^n W(\mu_k, \hat{\mu}_k) + n\epsilon. \end{aligned}$$

Letting $\epsilon \downarrow 0$ proves the result. $\square$

Lemma E.4 (Approximation error of Gumbel-Softmax layer). *Given $1 > \tau > 0$ and $p_1, \dots, p_n > 0$, let $\mu^\tau \sim X^\tau = (X_1^\tau, \dots, X_n^\tau)$ be*

$$X_k = \frac{\exp((\log p_k + G_k)/\tau)}{\sum_{k=1}^n \exp((\log p_k + G_k)/\tau)}, \quad (\text{E.3})$$

where $G_k \sim \exp(-x - \exp(-x))$ are i.i.d. standard Gumbel variables. Let μ^0 be the distribution of $X^0 = \lim_{\tau \rightarrow 0} X^\tau = (X_1^0, \dots, X_n^0)$. Then,

$$W(\mu^\tau, \mu^0) \leq 2(n-1)\tau - 2n(n-1)e^{-1}\tau \log \tau.$$

Remark. The gap event is $E_\delta = \{\exists 1 \leq i < j \leq n : |G_i + \log p_i - G_j - \log p_j| < \delta\}$, and its union bound counts each unordered pair once. In the displayed softmax formula, X_k means X_k^τ , $\mu^\tau = P(X^\tau)$, and $X^0 = \lim_{\tau \downarrow 0} X^\tau$ almost surely with the same i.i.d. standard Gumbel variables $G_1, \dots, G_n$ used for every τ .

Remark (notation). X^τ uses positive weights $p_1, \dots, p_n$, which need not sum to one, and $0 < \tau < 1$, with coordinates indexed by $k \in [n]$; all coordinates are positive and sum to one. In Theorem 3.6, the result split from Theorem 3.6: In the mixing construction, $n = N_C$ and $p_i = \mathbb{P}(C_i) > 0$ for $i \in [N_C]$; X^τ is the vector of Gumbel-softmax weights assigned to these support components. $X^0 = \lim_{\tau \downarrow 0} X^\tau = e_I$ almost surely, where $I = \arg \max_{k \in [n]} (\log p_k + G_k)$, using the same Gumbel draws throughout the limit; ties may be assigned to the first maximizing coordinate. In Theorem 3.6, the result split from Theorem 3.6: In the mixing construction, $n = N_C$ and $p_i = \mathbb{P}(C_i) > 0$ for $i \in [N_C]$; X^0 is the one-hot limit of these component weights, with $P(X^0 = e_i) = \mathbb{P}(C_i)$. μ^0 is the law of the one-hot vector X^0 , supported on the vertices of Δ_n , with $\mu^0(\{e_k\}) = p_k / \sum_{j=1}^n p_j$ for $k \in [n]$.

Proof. Without loss of generality, we may assume that $\sum_{k=1}^n p_k = 1$. Otherwise, we can divide the denominator and numerator in (E.3) by $\sum_{k=1}^n p_k$. Let $\Delta_n = \{(x_1, \dots, x_n) : \sum_{k=1}^n x_k = 1\}$. We construct a transportation map $T : \Delta_n \rightarrow \Delta_n$ as follows. $T(x_1, \dots, x_n)$ is a n -dimensional vector with all zeros except that the $i = \arg \max_j x_j$ -th entry being one. If there are multiple coordinates that is maximum, we choose the first coordinate and let the remaining coordinate be zero. It is easy to verify that $T_{\#}\mu^\tau$ and μ^0 have the same distribution [36]. For any $\delta > 0$, we have

$$\begin{aligned} P(|G_{k_1} + \log p_{k_1} - (G_{k_2} + \log p_{k_2})| < \delta) &= \int_{-\infty}^{\infty} \int_{y+\log \frac{p_{k_2}}{p_{k_1}}-\delta}^{y+\log \frac{p_{k_2}}{p_{k_1}}+\delta} \exp(-x - y - \exp(-y) - \exp(-x)) dx dy \\ &\leq 2\delta e^{-1} \int_{-\infty}^{\infty} \exp(-y - \exp(-y)) dy = 2\delta, \end{aligned}$$

where we have used $\exp(-x - \exp(-x)) \leq \exp(-1)$. Let event $E_\delta = \{\exists i, j \in [n] : |G_i + \log p_i - (G_j + \log p_j)| < \delta\}$, we have

$$\begin{aligned} P(E_\delta) &\leq \sum_{i \neq j} P(|G_i + \log p_i - (G_j + \log p_j)| < \delta) \\ &\leq \delta n(n-1)e^{-1}. \end{aligned}$$

Now, we calculate the transportation cost

$$\mathbb{E}_{X^\tau \sim \mu^\tau} [\|T(X^\tau) - X^\tau\|_1] \leq P(E_\delta) \mathbb{E}_{X^\tau \sim \mu^\tau} [\|T(X^\tau) - X^\tau\|_1 | E_\delta] + P(E_\delta^c) \mathbb{E}_{X^\tau \sim \mu^\tau} [\|T(X^\tau) - X^\tau\|_1 | E_\delta^c].$$

Note that $\|T(X^\tau) - X^\tau\|_1 \leq 2$. For the first term, we have

$$P(E_\delta) \mathbb{E}_{X^\tau \sim \mu^\tau} [\|T(X^\tau) - X^\tau\|_1 | E_\delta] \leq 2\delta n(n-1)e^{-1}.$$

For the second term, under event E_δ^c , let $k_{\max} = \arg \max_i X_i$, if $j = k_{\max} = \arg \max_i (\log p_k + G_k)$

$$\begin{aligned} \left| \frac{\exp((\log p_j + G_j)/\tau)}{\sum_{k=1}^n \exp((\log p_k + G_k)/\tau)} - 1 \right| &= \left| \frac{\sum_{k \neq j} \exp((\log p_k + G_k)/\tau)}{\sum_{k=1}^n \exp((\log p_k + G_k)/\tau)} \right| \\ &\leq \left| \frac{\sum_{k \neq j} \exp((\log p_k + G_k)/\tau)}{\exp((\log p_j + G_j)/\tau)} \right| \\ &= \sum_{k \neq j} \exp((\log p_k + G_k - (\log p_j + G_j))/\tau) \\ &\leq (n-1) \exp(-\delta/\tau). \end{aligned}$$

If $j \neq k_{\max}$, we have

$$\left| \frac{\exp((\log p_j + G_j)/\tau)}{\sum_{k=1}^n \exp((\log p_k + G_k)/\tau)} \right| \leq \exp((\log p_j + G_j - (\log p_{\max} + G_{\max}))/\tau) \leq \exp(-\delta/\tau).$$

Therefore,

$$\mathbb{E}_{X^\tau \sim \mu^\tau} [\|T(X^\tau) - X^\tau\|_1 | E_\delta^c] \leq 2(n-1) \exp(-\delta/\tau).$$

We get an upper bound of the transportation cost,

$$W(\mu^\tau, \mu^0) \leq \mathbb{E}_{X^\tau \sim \mu^\tau} [\|T(X^\tau) - X^\tau\|_1] \leq 2\delta n(n-1)e^{-1} + 2(n-1) \exp(-\delta/\tau).$$

Let $\delta = -\tau \log \tau$, we get

$$W(\mu^\tau, \mu^0) \leq 2(n-1)\tau - 2n(n-1)e^{-1}\tau \log \tau.$$

□

Lemma E.5 (Approximation Error of Sinkhorn distance). *For any $\mu \in \Delta_n$, $\nu \in \Delta_m$, and $\lambda > 0$, let μ and ν be the weights of finite probability measures on points $x_1, \dots, x_n$ and $y_1, \dots, y_m$ in a common Euclidean space with the paper's 1-norm. Put $c_{ij} = \|x_i - y_j\|$ and*

$$\Pi(\mu, \nu) = \left\{ T \in [0, \infty)^{n \times m} : \sum_j T_{ij} = \mu_i, \sum_i T_{ij} = \nu_j \right\}, \quad h(T) = 1 - \sum_{i,j} T_{ij} \log T_{ij},$$

with $0 \log 0 = 0$. Here the entropy regularized Wasserstein-1 distance is

$$W_{1,\lambda}(\mu, \nu) = \min_{T \in \Pi(\mu, \nu)} \left\{ \sum_{i,j} c_{ij} T_{ij} - \lambda h(T) \right\},$$

and $W(\mu, \nu) = \min_{T \in \Pi(\mu, \nu)} \sum_{i,j} c_{ij} T_{ij}$. Then

$$-(\log(mn) + 1)\lambda \leq W_{1,\lambda}(\mu, \nu) - W(\mu, \nu) \leq 0.$$

Moreover, we have the following estimation of approximation error.

$$|S_\lambda(\mu, \nu) - W(\mu, \nu)| \leq 2(\log(mn) + 1)\lambda.$$

Remark. The self terms in $S_\lambda(\mu, \nu) = W_{1,\lambda}(\mu, \nu) - \frac{1}{2}W_{1,\lambda}(\mu, \mu) - \frac{1}{2}W_{1,\lambda}(\nu, \nu)$ use the same 1-norm ground distance on the corresponding support points.

Remark (notation). $W_{1,\lambda}$ uses the supplied finite support points, the paper's 1-norm ground cost, and the entropy subtraction $-\lambda h(T)$ with $h(T) = 1 - \sum_{i,j} T_{ij} \log T_{ij}$ and $0 \log 0 = 0$.

Proof. The transportation polytope is nonempty and compact, and both objectives are continuous under $0 \log 0 = 0$, so their minima exist. Let T^* minimize the unregularized transport cost and let T_λ minimize the regularized objective. Since $h(T) \geq 0$ for every probability plan,

$$W_{1,\lambda}(\mu, \nu) \leq \sum_{i,j} c_{ij} T_{ij}^* - \lambda h(T^*) \leq W(\mu, \nu).$$

Also $\sum_{i,j} c_{ij} (T_\lambda)_{ij} \geq W(\mu, \nu)$. It follows that

$$-\lambda \max_{T \in \Pi(\mu, \nu)} h(T) \leq W_{1,\lambda}(\mu, \nu) - W(\mu, \nu) \leq 0. \quad (\text{E.4})$$

Since $T_{ij} \in [0, 1]$ and $f(x) = -x \log x$ is concave, we have

$$\begin{aligned} h(T) &= -nm \cdot \frac{1}{nm} \sum_{i,j=1}^{n,m} T_{ij} \log T_{ij} + 1. \\ &\leq -nm \cdot \left(\frac{1}{nm} \sum_{i,j=1}^{n,m} T_{ij} \right) \log \left(\frac{1}{nm} \sum_{i,j=1}^{n,m} T_{ij} \right) + 1 \\ &= 1 + \log(nm). \end{aligned}$$

By definition of Sinkhorn distance,

$$S_\lambda(\mu, \nu) = W_{1,\lambda}(\mu, \nu) - W_{1,\lambda}(\mu, \mu)/2 - W_{1,\lambda}(\nu, \nu)/2.$$

Let $a = W_{1,\lambda}(\mu, \nu) - W(\mu, \nu)$, $b = W_{1,\lambda}(\mu, \mu)$, and $c = W_{1,\lambda}(\nu, \nu)$. The self-transport Wasserstein costs are zero. By (E.4) and the entropy calculation,

$$|a| \leq (1 + \log(nm))\lambda, \quad |b| \leq (1 + 2 \log n)\lambda, \quad |c| \leq (1 + 2 \log m)\lambda.$$

Consequently

$$\begin{aligned} |S_\lambda(\mu, \nu) - W(\mu, \nu)| &= |a - b/2 - c/2| \\ &\leq |a| + |b|/2 + |c|/2 \\ &\leq 2(1 + \log(nm))\lambda. \end{aligned}$$

□

Lemma E.6. *Given a probability measure μ on $\mathbb{R}^{d'}$ and a class $\mathcal{F}$ of measurable functions $f : \mathbb{R}^{d'} \rightarrow \mathbb{R}^d$, suppose that each scalar coordinate class $\mathcal{F}_j = \{x \mapsto f_j(x) : f \in \mathcal{F}\}$, $j \in [d]$, has pseudo-dimension less than $d_{\mathcal{F}} < \infty$ and there exists $K > 0$ such that $\|f(x)\|_{\infty} \leq K$ for all $f \in \mathcal{F}$ and $x \in \mathbb{R}^{d'}$. Suppose in addition that there are a standard Borel space T and a jointly Borel map $G : T \times \mathbb{R}^{d'} \rightarrow [-K, K]^d$ such that $\mathcal{F} = \{G(t, \cdot) : t \in T\}$. There are positive constants C, c depending only on d, K such that, with probability at least*

$$1 - \exp \left(C \left((1 + \delta^{-d}) \log(2 + \delta^{-1}) + d_{\mathcal{F}} \log(2 + \epsilon^{-1}) \right) - cn\epsilon^2 \right),$$

$$\sup_{f \in \mathcal{F}} W(f_{\#}\mu, f_{\#}\mu_n) \leq \delta + \epsilon$$

for all $\delta, \epsilon > 0$, where μ_n is the empirical distribution of μ .

Remark. A network in $\mathcal{NN}_{k_1, k_2}(W, L)$ has L hidden layers, each of width at most W , followed by the affine output layer (width at most W and exactly W give the same class of functions: a narrower layer is padded with zero units). The covering number $\mathcal{N}(\epsilon, S, \|\cdot\|_1)$ is that of the cited source (Anthony and Bartlett): the least size of a subset $C \subseteq S$ such that every point of S lies at ℓ_1 distance strictly less than ϵ from some point of C . The probability bound is for each fixed pair $\delta, \epsilon > 0$. Probabilities of uniform deviation events are evaluated in the completion of the sample law. The boundedness condition means $\|f(x)\|_{\infty} \leq K$ for every $f \in \mathcal{F}$ and $x \in \mathbb{R}^{d'}$.

Remark (notation). $d_{\mathcal{F}} < \infty$ is a strict upper bound on the pseudo-dimensions of the scalar function classes specified in its local use. In Lemma E.6: For every $j \in [d]$, $\mathcal{F}_j = \{x \mapsto f_j(x) : f \in \mathcal{F}\}$ satisfies $\text{Pdim}(\mathcal{F}_j) < d_{\mathcal{F}}$; the bound applies separately to each coordinate class.

Proof. The empty class satisfies the uniform assertion vacuously, so assume $T \neq \emptyset$. Let $X_1, \dots, X_n$ be independent observations with law μ . By the dual formulation of Wasserstein distance,

$$\sup_{f \in \mathcal{F}} W(f_{\#}\mu, f_{\#}\mu_n) = \sup_{h \in \mathcal{H}} \sup_{f \in \mathcal{F}} (\mathbb{E}_{X \sim \mu_n} h(f(X)) - \mathbb{E}_{X \sim \mu} h(f(X))),$$

where

$$\mathcal{H} = \{h : [-K, K]^d \rightarrow \mathbb{R} : h(0) = 0, |h(x) - h(y)| \leq \|x - y\|\}.$$

Every such h takes values in $[-dK, dK]$. Retain the unnormalized empirical covering number

$$\mathcal{N}_1(r, \mathcal{F}, N) = \sup_{x_1, \dots, x_N} \mathcal{N}(r, \{(f(x_1), \dots, f(x_N)) : f \in \mathcal{F}\}, \|\cdot\|_1),$$

where the 1-norm on the evaluation vector sums over all sample points and output coordinates.

With the uniform norm, $\mathcal{H}$ is closed, uniformly bounded and equicontinuous on the compact cube, hence is a compact metric space by Arzela–Ascoli. Evaluation $(h, y) \mapsto h(y)$ is continuous: its increment is bounded by $\|h - h'\|_\infty + \|y - y'\|$. Consequently $(t, h, x) \mapsto h(G(t, x))$ is jointly Borel on $T \times \mathcal{H} \times \mathbb{R}^{d'}$. Its bounded population integral is Borel in (t, h) : this holds for indicators of measurable rectangles, is preserved by bounded monotone limits and linear combinations, and extends to bounded Borel functions by the monotone class theorem. Every strict deviation event is the projection of a Borel set from this standard Borel parameter space and is therefore analytic and universally measurable; weak supremum events are countable intersections of strict ones. The same reasoning applies to two samples and finite coordinate swaps. All probabilities below are thus defined in the completion of the sample law. Obviously, if h is 1-Lipschitz,

$$\mathcal{N}_1(\epsilon, h \circ \mathcal{F}, n) \leq \mathcal{N}_1(\epsilon, \mathcal{F}, n).$$

Put $v = \lfloor d_{\mathcal{F}} \rfloor$. Each coordinate class has pseudo-dimension at most v . Normalize its range by $f_j \mapsto (f_j + K)/(2K)$. Applying Theorem 18.4 of [2] in its averaged empirical metric, and then undoing the normalization, bounds a scalar unnormalized cover of radius $r/(2d)$ on N points by

$$e(v+1) \left(2 + \frac{8eKdN}{r} \right)^v.$$

The additional 2 allows radii beyond the normalized unit range. Take the product of these d covers. For every product ball meeting the evaluation vectors of $\mathcal{F}$, choose one such vector as its center. Two vectors in that product ball differ by at most r/d in each scalar unnormalized metric, hence by at most r in the summed vector metric. Thus

$$\mathcal{N}_1(r, \mathcal{F}, N) \leq \left[e(v+1) \left(2 + \frac{8eKdN}{r} \right)^v \right]^d.$$

This exhibits both sample-size and range scaling.

Fix $h \in \mathcal{H}$, and define

$$a_f(x) = \frac{h(f(x)) + dK}{2dK}, \quad z_f(x) = \sqrt{a_f(x)}.$$

Both functions take values in $[0, 1]$. On N points, the inequality $|\sqrt{a} - \sqrt{b}| \leq \sqrt{|a - b|}$ and the Cauchy–Schwarz inequality give

$$\frac{1}{N} \sum_{i=1}^N |z_f(x_i) - z_g(x_i)| \leq \left(\frac{1}{2dKN} \sum_{i=1}^N \|f(x_i) - g(x_i)\| \right)^{1/2}.$$

Consequently an unnormalized vector cover of radius $2dKNr^2$ gives an averaged scalar cover of the class $\{z_f : f \in \mathcal{F}\}$ of radius r , whose cardinality is at most

$$\left[e(v+1)(2 + 4e/r^2)^v \right]^d.$$

By Theorem 17.1 in [2], with its standing measurability condition, let X and T_0 be standard Borel spaces, let $g : T_0 \times X \rightarrow [0, 1]$ be jointly Borel, and put $\mathcal{G} = \{g(t, \cdot) : t \in T_0\} \neq \emptyset$. For every Borel probability distribution P on $X \times [0, 1]$, every $0 < \zeta < 1$ and every positive integer m ,

$$P^m \left\{ \text{some } t \in T_0 \text{ has } \left| \mathbb{E}_{(x,y) \sim P} (g(t, x) - y)^2 - \frac{1}{m} \sum_{i=1}^m (g(t, x_i) - y_i)^2 \right| \geq \zeta \right\} \leq 4\mathcal{N}_1^{\text{avg}}(\zeta/16, \mathcal{G}, 2m) \exp(-\zeta^2 m)$$

Here the draws are independent, the probability is taken in the completed sample law, and $\mathcal{N}_1^{\text{avg}}$ uses the averaged empirical metric. For the present class, $(t, x) \mapsto \sqrt{(h(G(t, x)) + dK)/(2dK)}$ is jointly Borel, since h is continuous and the square root is continuous on $[0, 1]$. Its parameter space is the same standard Borel space T , and its constant-response input law is $\mu \otimes \delta_0$. Apply this form of Theorem 17.1 to the class with response identically zero. Its squared loss is $z_f^2 = a_f$. For $0 < \epsilon < 2dK$, put $\eta = \epsilon/(2dK)$ and use the averaged cover at radius $\eta/16$ on $2n$ points. The theorem gives

$$\mathbb{P} \left(\sup_{f \in \mathcal{F}} |(\mu_n - \mu)(h \circ f)| > \epsilon \right) \leq 4 \left[e(v+1)(2 + 1024e/\eta^2)^v \right]^d \exp(-n\eta^2/32).$$

Since $v \leq d_{\mathcal{F}}$ and $\log(v+1) \leq v$, the logarithm of the right-hand side is bounded by $C_0 + C_1 d_{\mathcal{F}} \log(2 + \epsilon^{-1}) - c_1 n \epsilon^2$, with constants depending only on d, K . If $\epsilon \geq 2dK$, the deviation is deterministically at most $2dK$, and the desired deviation bound is immediate. Let $\mathcal{H}_\delta$ be a finite $\delta/2$ -net of $\mathcal{H}$ in the uniform norm. Such a net has

$$\log |\mathcal{H}_\delta| \leq C_2(1 + \delta^{-d}) \log(2 + \delta^{-1}),$$

where C_2 depends only on d, K . To see this, take a grid in $[-K, K]^d$ with 1-norm mesh at most $\delta/8$ and at most $(1 + C_3/\delta)^d$ points. Quantize the values of h at these points into bins of width at most $\delta/8$ in $[-dK, dK]$. There are at most $2 + C_4/\delta$ bins at each point. Choose one member of $\mathcal{H}$ from each nonempty quantization pattern. Two functions with the same pattern differ by at most $\delta/8$ at grid points and by at most $3\delta/8$ everywhere, using their Lipschitz bounds and a nearest grid point. These representatives form the required net. Counting patterns gives the displayed entropy bound. Apply the fixed- h concentration bound to every $h \in \mathcal{H}_\delta$ and take a union bound. For $0 < \epsilon < 2dK$, after enlarging a constant depending only on d, K ,

$$\mathbb{P} \left(\sup_{h \in \mathcal{H}_\delta, f \in \mathcal{F}} (\mu_n - \mu)(h \circ f) > \epsilon \right) \leq \exp \left(C \left((1 + \delta^{-d}) \log(2 + \delta^{-1}) + d_{\mathcal{F}} \log(2 + \epsilon^{-1}) \right) - cn\epsilon^2 \right).$$

For every $h \in \mathcal{H}$, choose $\tilde{h} \in \mathcal{H}_\delta$ with $\|h - \tilde{h}\|_\infty \leq \delta/2$. Since μ_n and μ are probability measures,

$$|(\mu_n - \mu)((h - \tilde{h}) \circ f)| \leq \delta.$$

Therefore

$$\sup_{h \in \mathcal{H}, f \in \mathcal{F}} (\mu_n - \mu)(h \circ f) \leq \delta + \sup_{\tilde{h} \in \mathcal{H}_\delta, f \in \mathcal{F}} (\mu_n - \mu)(\tilde{h} \circ f).$$

The dual formula yields the stated Wasserstein bound. If $\epsilon \geq 2dK$, both push-forward measures are supported in a set of 1-norm diameter $2dK$, so their Wasserstein distance is at most $2dK \leq \delta + \epsilon$ deterministically. This completes the proof for all positive δ, ϵ . $\square$

By Theorem 18.4 in [2], if $\mathcal{G}$ is a nonempty class of functions from a set X into $[0, 1]$ with finite pseudo-dimension v , then for every $0 < \epsilon \leq 1$ and every positive integer m ,

$$\mathcal{N}_1^{\text{avg}}(\epsilon, \mathcal{G}, m) \leq \mathcal{M}_1^{\text{avg}}(\epsilon, \mathcal{G}, m) < e(v+1) \left(\frac{2e}{\epsilon} \right)^v,$$

where $\mathcal{N}_1^{\text{avg}}(\epsilon, \mathcal{G}, m)$ and $\mathcal{M}_1^{\text{avg}}(\epsilon, \mathcal{G}, m)$ are the largest, over $x_1, \dots, x_m \in X$, covering and packing numbers of $\{(g(x_1), \dots, g(x_m)) : g \in \mathcal{G}\}$ in the averaged metric $d_1(u, w) = \frac{1}{m} \sum_{i=1}^m |u_i - w_i|$.

This is Theorem 18.4 in [2].

The cited assertion used in Lemma E.6 is the following restricted form of Theorem 17.1 in [2], with its standing measurability condition. Let X and T be standard Borel spaces, let

$g : T \times X \rightarrow [0, 1]$ be jointly Borel, and put $\mathcal{G} = \{g(t, \cdot) : t \in T\} \neq \emptyset$. For every Borel probability distribution P on $X \times [0, 1]$, every $0 < \epsilon < 1$ and every positive integer m ,

$$P^m \left\{ \text{some } t \in T \text{ has } \left| \mathbb{E}_{(x,y) \sim P} (g(t, x) - y)^2 - \frac{1}{m} \sum_{i=1}^m (g(t, x_i) - y_i)^2 \right| \geq \epsilon \right\} \\ \leq 4\mathcal{N}_1^{\text{avg}}(\epsilon/16, \mathcal{G}, 2m) \exp(-\epsilon^2 m/32).$$

Here the sample consists of independent draws from P , the probability is taken in its completion, and $\mathcal{N}_1^{\text{avg}}(r, \mathcal{G}, N)$ is the supremum over $x_1, \dots, x_N \in X$ of the covering number of the evaluation vectors in the metric $N^{-1} \sum_i |u_i - v_i|$. This structural domain supplies the standing measurability condition: the loss is bounded and jointly Borel, and its integral against P is Borel in t . The latter assertion holds for indicators of measurable rectangles, is preserved by bounded monotone limits and linear combinations, and hence holds for bounded Borel functions by the monotone class theorem. Each bad event is a projection of a Borel subset of $T \times (X \times [0, 1])^m$, so it is analytic and universally measurable. Replacing the sample by two independent samples or applying any of the finitely many coordinate swaps gives Borel functions on finite products and therefore the same measurability conclusion. Thus the sample and ghost-sample events used by Theorem 17.1 are measurable in the completed sample laws, and that theorem gives the displayed inequality.

Remark (notation). $\mathcal{N}_1(\epsilon, \mathcal{F}, n) = \sup_{x_1, \dots, x_n} \mathcal{N}(\epsilon, \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\}, \|\cdot\|_1)$, where the points range over the domain of $\mathcal{F}$ and the norm is the printed ℓ_1 norm without a factor $1/n$. In a result cited in the proof of Lemma E.6: $\mathcal{N}_1^{\text{avg}}(\epsilon, \mathcal{G}, m)$ is the supremum over $x_1, \dots, x_m \in X$ of the covering numbers of $\{(g(x_1), \dots, g(x_m)) : g \in \mathcal{G}\}$ in $d_1(u, w) = m^{-1} \sum_{i=1}^m |u_i - w_i|$, for $\epsilon > 0$ and positive integers m . This averaged covering number is distinct from the covering number defined with the unnormalized ℓ_1 norm. In a result cited in the proof of Lemma E.6: $\mathcal{N}_1^{\text{avg}}(r, \mathcal{G}, N)$ is the supremum over $x_1, \dots, x_N \in X$ of the covering number of $\{(g(x_1), \dots, g(x_N)) : g \in \mathcal{G}\}$ in $d_1(u, w) = N^{-1} \sum_{i=1}^N |u_i - w_i|$. The probability bound uses $r = \epsilon/16$ and $N = 2m$. This averaged covering number is distinct from the covering number defined with the unnormalized ℓ_1 norm. $\mathcal{H}_\delta \subseteq \mathcal{H}$ is a finite $\delta/2$ -net in the uniform norm of $\mathcal{H} = \{h : [-K, K]^d \rightarrow \mathbb{R} : h(0) = 0, |h(x) - h(y)| \leq \|x - y\|\}$, for $\delta > 0$.