A TOPOLOGY-GUIDED CORRUPTION PROCESS FOR DISCRETE DIFFUSION ON TABULAR DATA

Emily Cheng, Aymane El Gadarri, Adam Jozefiak

[1]MIT EECS [2]MIT ORC [3]MIT ORC

{emilych, aymelgad, jozefiak}@mit.edu

ABSTRACT

Detailed household-level energy consumption data are essential for evaluating the distributional consequences of climate regulations, yet existing public surveys, such as the U.S. Residential Energy Consumption Survey (RECS), provide too few observations per census tract to support tract-scale analysis. We address this gap by developing a diffusion-based generative model that produces realistic synthetic household data conditioned on tract-level census statistics. Our approach tackles two key challenges in this setting: (1) mixed continuous and discrete features, and (2) strong hierarchical dependencies among variables. To handle discrete data, we build upon recent advancements in discrete diffusion modeling, particularly methods such as TabDDPM [1] and TabDiff [2], which discretize the diffusion process through categorical noise transition matrices, effectively extending diffusion methods to discrete tabular domains. To address hierarchical dependence, our main contributions include (i) a **structure-aware noise schedule** that injects noise from leaves to root along an approximate Chow–Liu tree constructed from the variables, and (ii) a **masked self-attention denoiser** whose receptive field aligns with the same graphical structure. Through extensive experiments, we show that sequential diffusion models beat standard ones on tree-like graphical models by imposing an inductive bias on the noise schedules. On the other hand, we find that their performance degrades gracefully with data that is only approximately tree-like, performing slightly better than standard diffusion on the RECS dataset, while also improving a lot on generating truly tree-like graphical models. This paves the way to future work that can further balance the tradeoff between the approximation and the estimation errors.

1. INTRODUCTION

Our project explores the application of diffusion-based generative models, specifically to sample from mixed distributions that have both numerical and categorical variables. A particular dataset of interest that we set as the ultimate benchmark for this type of generation is household energy consumption data. We aim beyond this project to address the challenge faced by policymakers who currently lack sufficiently granular data for accurately evaluating climate regulation and tax policies' impacts at the household level. By generating realistic synthetic household-level data, we aim to inform equitable and effective policymaking in energy and climate policy.

More specifically, policymakers often require detailed household consumption data to assess the distributional impacts of climate regulations, such as carbon pricing and energy subsidies. However, available survey datasets (e.g., Residential Energy Consumption Survey - RECS) are limited in scope, containing insufficient observations per census tract. The core problem this project addresses is: Given aggregate census tract-level statistics and limited household-level survey data, how can we generate realistic household consumption profiles conditioned on tract-level statistics using diffusion models? To answer this question, conditional generation is straightforward. The demographic data present in RECS is so extensive, as detailed in the next section, that knowing it fully is enough to generate the electricity consumption variables. In other words, we can fairly assume that conditioned on the demographic data, the energy consumption data is independent of the census tract. This implies that in order to generate conditionally from the census tract, we can first sample a household's demographic data from that tract, and then sample energy consumption data conditionally from the demographics. The crux of our problem is thus reduced to learning a diffusion model on the joint distributions of demographics and energy consumption.

The key challenge then arises from the structure of the RECS dataset, which contains both continuous and discrete features. While diffusion models for continuous data are well studied, their application to mixed-type tabular data remains nascent. Recent methods like TabDDPM and TabDiff [1, 2] extend diffusion models to tabular domains, but these approaches are not yet fully mature. Our goal is to advance these models by improving sample fidelity and computational efficiency, using our policy-driven application as a natural testbed.

Tabular datasets like RECS are often highly structured and hierarchical. For example, the climate region variable influences heating and cooling degree days, which—together with home size—correlate with energy usage for heating and cooling, a major component of total energy consumption. This structure motivates us to incorporate an approximate graphical model into both the noise schedule and the model architecture. Specifically, we use a Chow–Liu tree to approximate the joint distribution and design a noise schedule that adds noise progressively from leaves to root, leveraging causal structure similar to diffusion forcing and independent noise levels introduced in [**?**]. Tangential to designing a structurally informed noise schedule, we can develop diffusion matrices for the discrete variables that are likewise informed by the distribution's graphical structure. On the architectural side, we propose to apply masked self-attention aligned with a given graphical model, following the approach of [3, 4] which showed that masking according to a known graphical model yields superior quality of samples and training and inference speedup. We propose to learn an approximate graphical model of the data using the classical Chow-Liu decomposition [5].

A core part of our work is to try to understand the limitations of current diffusion models when dealing with highly structured and correlated distributions. We design a distribution that current diffusion models struggle a lot with, but that graphical models can learn fairly well.

We find that our topology-guided diffusion model performs well on such tree-like distributions. Beating easily the standard diffusion model that cannot disentangle the highly correlated tree model.

On the other part, the performance of sequential diffusion degrades gracefully with learning distributions that are only approximately tree-like. We find that it performs slightly better than standard diffusion on the RECS dataset that is of interset.

This suggests a fundamental trade-off between the approximation and estimation errors, and opens future research investigations.

2. RELATED WORK

Discrete Diffusion for Tabular Data. Diffusion models have demonstrated remarkable success in continuous domains (e.g., images and audio), but their extension to mixed-type tabular data poses unique challenges due to the presence of discrete and often high-cardinality categorical features. TabDDPM [1] was among the first to tackle this by formulating a denoising diffusion probabilistic model (DDPM) over discrete variables via learnable categorical noise transition matrices. By replacing the Gaussian perturbation in standard DDPMs with a Markov chain over categories, TabDDPM enables end-to-end training on purely discrete tables, but can suffer from mode collapse when faced with complex inter-feature dependencies. Building on TabDDPM, TabDiff [2] introduces a multi-modal framework that jointly models continuous and discrete columns: continuous features are diffused in the usual Gaussian sense, while discrete features use parameterized transition kernels. TabDiff further stabilizes training by conditioning the denoiser on a reference noise schedule, improving sample quality and convergence compared to TabDDPM. However, both approaches rely on uniform noise schedules and dense self-attention, which may struggle to capture strong hierarchical or tree-like dependencies present in richly structured datasets such as RECS.

Graph-Structured Diffusion Models. Graphically Structured Diffusion Models (GSDMs) [3, 4] incorporate known graphical structure directly into the diffusion process and model architecture. Weilbach et al. [3] show that by masking self-attention according to the adjacency of a given graph, one can reduce computational complexity from $O(d^2)$ to $O(|E|)$ per layer and improve sample fidelity on problems with explicit combinatorial structure. In follow-up work [4], the authors scale this approach to larger graphs and demonstrate that sparse attention not only accelerates training but also mitigates mode collapse in settings where the true distribution has a clear dependency graph. While GSDMs assume the graph is known a priori, their masking and noise-ordering principles suggest a powerful inductive bias for any strongly structured tabular domain.

3. DATA

3.1. Real-world datasets

3.1.1. Residential Energy Consumption Survey (RECS)

We've so far hinted at the challenging nature of the RECS dataset, with both numerical and categorical features and highly dependent features. The Residential Energy Consumption Survey (RECS), administered by the U.S. Energy Information Administration (EIA) through web and mail questionnaires, provides a nationally representative sample of housing units across the country. Combined with data from energy suppliers, it captures information on household energy characteristics and is designed to help inform future projections of energy demand. Given its national scope and rich feature set, the RECS dataset has been widely used in research on residential energy use. Prior studies have aimed to analyze household energy consumption patterns and their determinants, identify gaps in energy efficiency and policy implementation, and develop predictive methods for residential energy performance.

The RECS dataset boasts 18496 rows and 786 features. 61 of such features are of category weight, which are for survey estimation and variance calculations. Of the remaining features, 338 are categorical, while 387 are numerical, which we found by manually combing through the dataset. The features are categorized into one of the following [1]:

[1] There is another category for Weights, as well as one titled End-use Model. End-use Model contains calibrated and total energy use and cost estimates across fuels and appliances, based on billing data and modeled consumption patterns.

- **Admin** (4): Includes climate zone classifications, urbanization type, and unique household identifiers for organizing the dataset.
- **Air Conditioning** (40): Captures details on air conditioning equipment, usage patterns, cooling-related appliances, and associated imputation indicators.
- **Appliances** (108): Includes detailed information on kitchen and laundry appliances, their characteristics, fuel types, usage frequency, and imputation indicators.
- **Electronics** (96): Covers television characteristics, teleworking devices, smart home technologies, internet access, and associated usage and imputation indicators.
- **Energy Assistance** (46): Captures household experiences with energy-related financial hardship, disconnection, and participation in assistance programs.
- **Energy Bills** (33): Details how households pay for and interact with energy utilities, covering fuel payment responsibilities, smart meter access, electric vehicle charging, outages, and derived energy use types.
- **Geography** (5): Categorizes households by Census region and division and provides state-level identifiers.
- **Household Characteristics** (24): Captures respondent demographics (e.g. employment, education, race, age, sex), household composition, income, weekday presence at home, and imputation flags for each.
- **Lighting** (22): Tracks the types, usage durations, and nighttime behavior of light bulbs (LED, CFL, incandescent/halogen) inside and outside the home, along with corresponding imputation flags.
- **Space Heating** (43): Details heating system types, fuels, usage, and locations, plus humidifier info and imputation flags.
- **Thermostat** (17): Captures thermostat type, temperature control habits (manual, programmable, smart), seasonal settings (summer/winter, home/away/night), and imputation flags.
- **Water Heating** (22): Covers water heater characteristics (size, location, age,), fuel types (primary vs. secondary), equipment count, energy source (electricity, gas, solar, etc.), and related imputation flags.
- **Weather** (4): Includes measures of annual heating and cooling needs based on local climate data.
- **Your Home** (89): Describes structural and ownership characteristics of the housing unit, including foundation, insulation, room counts, square footage, and windows, and includes both respondent-reported and derived variables, along with imputation flags.

Admittedly, with imputed values, we risk systematic bias, and variables with high imputation rates might not reflect real household behavior. With surveys, there may be gaps from underrepresented populations and noise in variables that require more precise estimates.

3.1.2. Classical Benchmarks for Tabular Diffusion

While the authors of TABDIFF claim that their seven datasets—rAdult, Default, Shoppers, Magic, Beijing, News, and Diabetes—are representative of real-world data, we argue that these benchmarks remain overly simplistic compared to truly complex real-world datasets such as RECS. An examination of their high-level properties in Table 2, reveals significant disparities that underscore the increased challenge of RECS.

Table 1. Overview of dataset's statistics

Dataset	# Rows	# Num	# Cat	# Max Cat
Adult	48,842	6	9	42
Default	30,000	14	11	11
Shoppers	12,330	10	8	20
Magic	19,019	10	1	2
Beijing	43,824	7	5	31
News	39,644	46	2	7
Diabetes	101,766	9	27	716
RECS	18,496	387	338	51

- **Number of Rows**: TABDIFF datasets provide a range of sample sizes, from 12,000 to over 100,000, and the RECS dataset contains 18,496.
- **Number of Numerical Features**: The TABDIFF datasets contain a modest number of numerical features, ranging from 6 to 46; the RECS dataset, however, is comprised of 387 numerical features.
- **Number of Categorical Features**: The number of categorical features in the TABDIFF datasets ranges from 1 to 27. RECS includes 338 categorical features.
- **Maximum Categories**: While the Diabetes dataset has a high # Max Cat value of 716 for one of its 27 categorical columns, and Adult has 42 for one of its 9, RECS presents a different type of categorical challenge. With 338 distinct categorical features, even with a # Max Cat of 51 (which represents the number of states) the combinatorial complexity across these numerous features is immense.

The significantly higher feature count to those of TABDIFF datasets means that learning the underlying data distribution and interfeature dependencies is an inherently more difficult task. Thus, generative models for RECS must capture a far richer and more intricate set of relationships.

3.2. Synthetic datasets

As mentioned in the introduction, we design a synthetic distribution that current diffusion models struggle a lot with. The goal is to elicit one of the root causes that hinder the performance of the current paradigms.

Let $(X_1, \ldots, X_n)$ be a set of n random variables in $\mathbb{R}^2$. The set follows a binary tree model called spiral model, where the depth of the layer corresponding to node k is $\lceil \frac{k}{2} \rceil$. Let $X_0 \sim \mathcal{N}(0, \sigma_0 I_2)$. For a node $X_k, k \geq 1$, let X_l, X_r be its left and right children that are in layer d. Then for $i \in l, r$, X_i follows the conditional distribution $X_i \sim \mathcal{N}(R_{\theta_i^d} X_k + t^i, \sigma I_2)$.

Where:

- R_θ is the rotation matrix with angle θ.
- $\theta_r^d = (-1)^{\delta_{i=r}} \alpha 2^d \mod 2\pi$
- $t^i = ((-1)^{\delta_{i=r}} \delta, 0)$
- D is the depth of the tree.

The parameters of this tree are thus $\{\sigma_0, \sigma, \alpha, \delta\}$. The tree-structure of the model is shown in figure 1. Intuitively, we embed a binary tree of 2-D nodes by mapping each parent to its children via a depth-doubling rotation $R_{\pm \alpha 2^d}$, a small translation $\pm \delta \mathbf{e}_x$, and isotropic Gaussian noise $\mathcal{N}(0, \sigma^2 I)$. The parameters α (base angle), δ (sibling offset), σ (manifold thickness), and depth D control, respectively, the folding frequency, branch overlap, ribbon width, and the exponential number (2^{D-1}) of terminal modes. Because each edge-wise conditional remains a unimodal 2-D Gaussian, a graph-aware model can learn the exact score, whereas a standard diffusion model must approximate a globally singular score field that flips direction every $O(\sigma)$ in $\mathbb{R}^{2D}$, leading to mode collapse, which commonly refers to the phenomenon where a model assigns too little probability to some modes of the true distribution. We find in our experiments, both in highly correlated real-world datasets, and for this synthetic distribution, that this mode-collapse is particularly exacerbated. This phenomenon is highlighted in previous work [6] [7].

4. MODELS AND METHODS

We utilize the TabDiff model [2] as a baseline model for tabular diffusion as it has been shown to produce higher quality samples and more stable training than its preceding competitor TabDDPM [1]. We train TabDiff on the RECS dataset with optimal hyperparameters and compare it on key metrics to our model defined below.

4.1. Chow–Liu Tree Approximation

In order to exploit the hierarchical and causal structure of the RECS dataset—and more broadly of mixed-type tabular data—we use the Chow–Liu algorithm to obtain an approximate, tree-structured graphical model of our joint distribution. The Chow–Liu dependence structure fits a multivariate distribution with a tree by minimizing the Kullback–Leibler divergence to the true distribution [5].

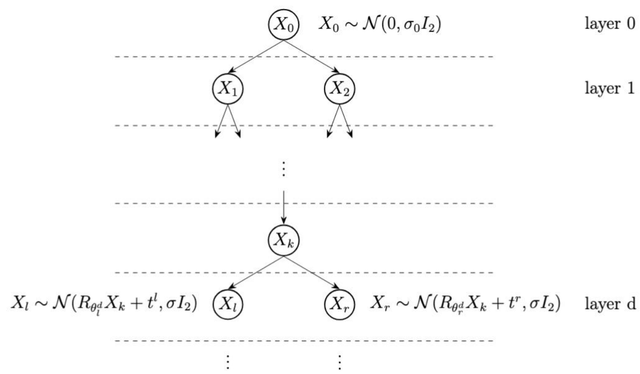

Fig. 1. The binary tree structure of the spiral model

Formally, let $\mathbf{X} = (X_1, \ldots, X_d)$ have true joint density $p(\mathbf{x})$. We seek a tree-structured distribution

$$q^* = \arg\min_{q \in \mathcal{T}} D_{\mathrm{KL}}(p \,\|\, q),$$

where $\mathcal{T}$ is the set of all tree-factorized distributions. Chow and Liu show this is equivalent to finding the maximum-spanning tree on the complete graph whose edge weights are the pairwise mutual informations $I(X_i; X_j)$. Importantly, these mutual information terms are tractable to estimate—using simple histogram or kernel-based methods—and can be computed once from the data prior to training the diffusion model.

The resulting tree T induces the factorization

$$p_{\mathrm{CL}}(\mathbf{x}) = p(x_r) \prod_{(i \to j) \in T} p(x_j \mid x_i),$$

with r an arbitrarily chosen root. By capturing the strongest dependencies, this approximation remains both expressive and tractable.

In our work, we leverage the Chow–Liu tree to

1. Order the diffusion noise schedule—adding noise progressively from the leaves inward, and
2. Impose a masked self-attention pattern in our model architecture that mirrors the dominant conditional dependencies in the data.

There may be other approaches of generating approximate graphical representations of the joint distribution using tractable computations such as the marginal and pairwise joint marginal distributions, which are present in the Chow-Liu algorithm. For instance, we could sparsify the dense graphical model according to the mutual information matrix using spectral and $(1+\epsilon)$-cut sparsifiers, but we leave this for future work.

4.2. Structurally Defined Noising

To incorporate the structural information captured by the Chow–Liu tree, we define a diffusion noise schedule that adds noise progressively from the leaves toward the root. This mirrors the direction of conditional dependence in the tree and aligns with ideas from diffusion forcing and independent noise scheduling in [8], which leverage causal structures in sequential data—often motivated by autoregressive language models. While tabular data lacks an obvious temporal or causal order, the Chow–Liu tree provides a principled approximation of the dependency structure, which we use to guide the denoising process in the reverse direction: from the root to the leaves. However, this structure-aware denoising may require additional steps, since leaf nodes are denoised more gradually or with a delay compared to their ancestors. The key empirical question is whether this structure-informed schedule improves sample quality without incurring significant computational overhead—or, ideally, even reduces it.

4.3. Graph-Structured Self-Attention Masking

Following the Graphically Structured Diffusion Models (GSDM) of Weilbach et al. [3], we impose a sparse attention mask based on the Chow–Liu tree in every self-attention layer of our transformer backbone. Let T be the set of undirected edges in the Chow–Liu tree over $\{1, \ldots, d\}$. We construct a binary mask $M \in \{0,1\}^{d\times d}$ by

$$M_{ij} = \begin{cases} 1, & \text{if } (i,j) \in T \text{ or } i = j, \\ 0, & \text{otherwise.} \end{cases}$$

Then each attention layer computes

$$\mathrm{Attention}(Q, K, V) \;=\; \mathrm{softmax}\big(M \odot (QK^T)\big)\, V\,,$$

so that tokens attend only to their tree-neighbors. This reduces per-layer complexity from $O(d^2)$ to $O(E)$, where $E = |T|$, yielding substantial speedups. Empirically, this structure-aware masking improves sample fidelity and lowers both training and inference costs [3, 4].

While sparse self-attention has been introduced by Weilbach et al. for combinatorial problems with explicit graphical models, our novelty lies in deriving the tree structure automatically via the Chow–Liu approximation on tabular data. Moreover, we investigate more liberal autoregressive masking schemes, where each node attends to all to all nodes in preceding tree layers. We find that this masking scheme enjoys better performance, since it is better to condition on all previously fully unmasked variables when the tree model is only approximate.

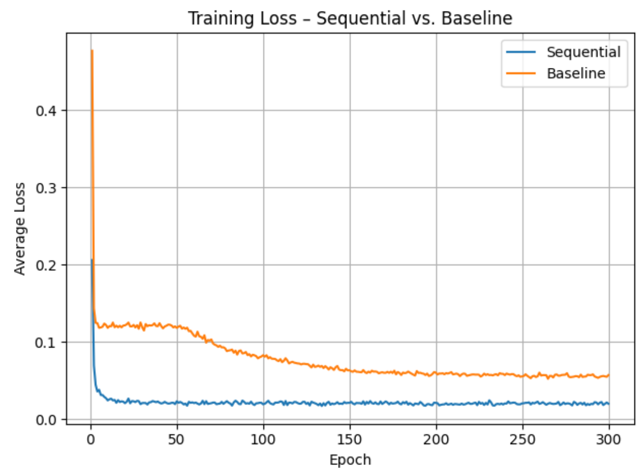

Fig. 2. Training loss of the denoiser network on the spiral distribution. Blue: Our sequential model training loss. Orange: The baseline model.

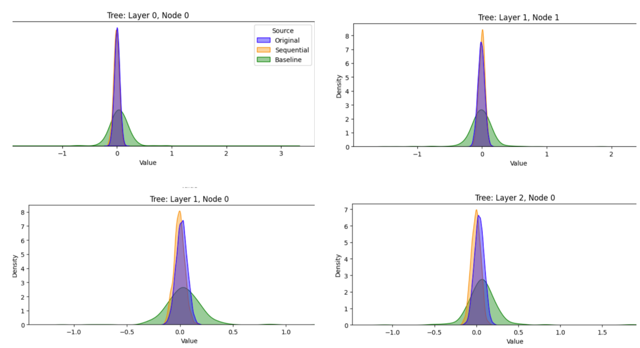

Fig. 3. Marginal densities for the first 4 nodes x coordinates for the spiral distribution. Purple: True distribution, Orange: Our sequential model's distribution, Green: Vanilla TABDiff.

5. EXPERIMENTS

5.1. Experiments on synthetic data

We train both a base DDPM model and our sequential diffusion model on the synthetic spiral distribution. Both models have the exact same architecture, and differ only through the noising and denoising schedules. We follow a variance preserving DDPM formulation. The denoising network takes a noisy variable x_t, and embeds each of its coordinates into 4 dimensions, as is standard. The time step is then embedded with a sinusoidal positional embedding and passed through a small MLP(2 layer with GeLu activation and hidden dimension of 128), before being concatenated to the hidden representation of the noisy sample in the first layer of the MLP (512 hidden dimension). We train both models with optimal hyperparameters and record the training loss in figure 2.

We see that the loss converges to a lower minima for the sequential model. Furthermore, the marginal density plots of the random variables confirm that the baseline completely fails to cover the modes of the true distribution. This is an instantiation of the mode collapse problem highlighted previously in the litterature. The sequential model on the other hand has an easier task learning the highly correlated distributions given the exact graphical structure of the model. Where the sequential model light fail is when the graphical model learned is only approximate. This suggests an estimation approximation errors tradeoff that our model is going to face in the next section.

Sequential diffusion can be seen as imposing an implicit bias on the noise schedules to allow effective learning. In fact, one of the pitfalls of current diffusion paradigms is that learning noise schedule parameters naively is hard [9]. Current work, such as [2] tries to mitigate the problem by facilitating learning the gradient of the diffusion loss with regards to the noise parameters by conditioning the denoising network, on top of the current noisy estimate x_t, with a known good enough noise schedule with fixed parameters $\sigma_t^{\theta_0}$, rather than the actual noise schedule. This makes the gradient of the noise parameter only go through its presence in calculating the noisy sample x_t, and not through the encoding of the noise level itself. In fact, we notice in our experiments that for the original model, the noise schedule parameters get stalled and don't change much as training progresses. This makes the case for a need to include an additional inductive bias to the noise schedules.

5.2. RECS Experiments

On the task of sampling from the RECS dataset, we evaluated TabDiff and TabDiff-Seq, a modification of the TabDiff model that incorporates a sequential noise schedule according to a graphical model capturing the dependencies of the dataset. Given the high-dimensional nature of the RECS dataset we limited our experiments to 75 variables, 200 training epochs, and 90 denoising steps. An initial Chow-Liu tree-approximation of the 75-variable subset of the RECS dataset was performed, illustrated by Figure 4.

A hyperparameter sweep was performed, centered around the optimal hyperparameters presented in the TabDiff paper [2]. Ultimately, for both TabDiff and TabDiff-Seq we selected the same hyperparameters as in the TabDiff paper with the exception of a learning rate of 10^{-4} and $\sigma_{\text{data}} = 1.5$. We found that TabDiff-Seq attains superior sample-quality scores, shape and trend, relative to TabDiff, see Table 2. The shape score measures how similar the marginals of the generated synthetic data are to the marginals of the true dataset. While the trend score measures how similar the pair correlations of the generated synthetic data are to the correlations of the true dataset. In particular, the shape is the Kolmogorov–Smirnov statistic for numerical columns and

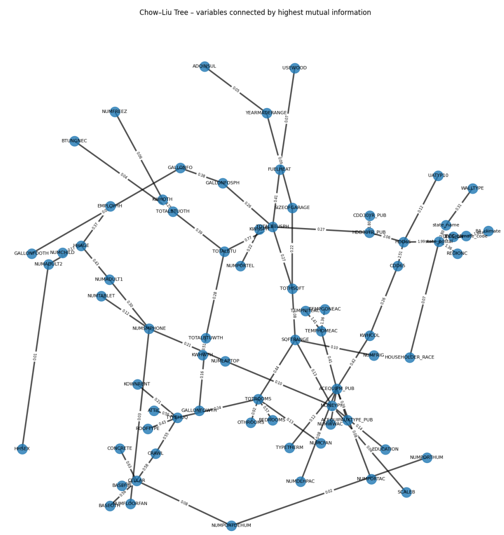

Fig. 4. Chow-Liu tree approximation of a 75-variable subsets of the RECS.

is the total variation distance for the categorical variables. While the trend is the average absolute error in Pearson correlations for numerical-numerical pairs and is the total variation distance for categorical-categorical pairs.

Table 2. Synthetic Sample Quality for RECS - Sequential vs. Non-Sequential Noise Schedules. Larger score means better approximation of the distribution by the synthetic samples.

Dataset	Shape	Trend	Overall
TabDiff-Seq.	0.879	0.852	0.865
TabDiff	0.852	0.811	0.836

When implementing TabDiff-Seq, we found in our experiments, that directly passing the timestep embedding rather than the noise level, as is implemetned in TabDiff, yields higher quality samples (in terms of shape and trend scores). This architectural modification is motivated by the fact that under our sequential noising the total noise is no longer an injective mapping outside the interval corresponding to the layer being. We additionaly considered adding an explicit layer embedding on top of the time embedding, in order to further encode the graphical structure of the Chow-Liu tree-approximation, however our experiments found that this architectureal change degraded performance.

6. CONCLUSION AND FUTURE WORK

Following the success of diffusion models in generating realistic continuous data, extending these methods to discrete data has become an exciting and fast-growing research area.

Yet discrete data remain challenging: their combinatorial nature and the absence of gradient signals make them harder to model than their continuous counterparts.

In this work, we first argue that the benchmarks most commonly used today—already solved by state-of-the-art (SOTA) models—are no longer adequate because they contain only a limited number of categories. As a more demanding benchmark, we propose to examine the Residential Energy Consumption Survey (RECS), whose highly correlated and structured variables have not yet been captured by any generative model, including mixed-data diffusion models such as TABDiff.

We then introduce a synthetic distribution purpose-built to expose diffusion models' failure modes, showing that they cannot learn a highly correlated binary tree. To overcome this weakness, we present a sequential diffusion model that learns an approximate tree structure and uses it to bias the noise schedule. Conceptually, the model bridges autoregressive and diffusion paradigms by fully denoising the data layer by layer.

Our experiments demonstrate that, on the synthetic datasets, our sequential model approximates the true distribution far better than standard diffusion baselines, which fail completely in this highly structured setting. On RECS, it slightly outperforms standard TABDiff. These findings suggest that graphical-model induced inductive biases are a promising avenue for discrete distributions, although the optimal form of the bias remains an open question.

As future work, we plan to investigate alternative denoising schemes—such as soft sequential models with overlapping denoising schedules—to balance estimation and approximation errors when the underlying distribution is only approximately tree-like.

7. REFERENCES

[1] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko, "Tabddpm: Modelling tabular data with diffusion models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 17564–17579.

[2] Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec, "Tabdiff: a multimodal diffusion model for tabular data generation," *arXiv preprint arXiv:2410.20626*, 2024.

[3] Christian Dietrich Weilbach, William Harvey, and Frank Wood, "Graphically structured diffusion models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 36887–36909.

[4] Christian Dietrich Weilbach, William Harvey, Hamed Shirzad, and Frank Wood, "Scaling graphically structured diffusion models," in *ICML 2023 Workshop on Structured Probabilistic Inference* {\&} *Generative Modeling*, 2023.

[5] CKCN Chow and Cong Liu, "Approximating discrete probability distributions with dependence trees," *IEEE transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[6] Sumukh K Aithal, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter, "Understanding hallucinations in diffusion models through mode interpolation," 2024.

[7] Roberto Barceló, Cristóbal Alcázar, and Felipe Tobar, "Avoiding mode collapse in diffusion models fine-tuned with reinforcement learning," 2024.

[8] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann, "Diffusion forcing: Next-token prediction meets full-sequence diffusion," 2024.

[9] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang, "Common diffusion noise schedules and sample steps are flawed," 2024.