Generative Causal Models: A Theoretical Framework for Integrating Flow Matching and Causal Inference

Kweku A. Opoku-Agyemang*

July 11, 2025

Abstract

This paper introduces a novel theoretical framework for generative causal models, which integrates Flow Matching, a state-of-the-art generative modeling technique, with causal inference to address the challenge of generating data that inherently respects causal structures. We propose a mathematical formulation where Flow Matching, defined by a vector field that transports noise to data along a specified path, is constrained by causal graphs to ensure that the generated samples adhere to causal relationships. Specifically, we consider a directed acyclic graph, representing variables and their causal dependencies, and define a generative process consistent with this graph. The Flow Matching objective is adapted to minimize the discrepancy between the learned vector field and a target vector field derived directly from the causal structure. We prove that under certain regularity conditions, the optimized vector field converges to this target, ensuring that the generated samples are causally consistent. Furthermore, we extend this framework to handle interventional and counterfactual scenarios by defining conditional Flow Matching objectives that respect do-calculus operations. For an intervention, the generative process is modified to align with the post-interventional distribution, and we derive bounds on the approximation error in terms of the causal graph's structure and the Flow Matching model's capacity. This theoretical advancement not only enhances the interpretability and robustness of generative models but also provides a new tool for causal discovery and policy analysis in high-dimensional settings. Our results suggest that generative causal models can significantly improve the estimation of causal effects in complex systems, offering a unified approach to generation and inference that is both theoretically grounded and practically scalable.

*Machine Learning X Doing and International Growth Centre. Email: kweku@machinelearningxdoing.com. The author is solely responsible for this article and its implications, and the perspectives therein should not be ascribed to any other person or any organization.

1 Introduction

The pursuit of robust causal inference is central to scientific discovery and policy formulation across myriad disciplines, none more so than economics. Understanding "what if" scenarios—the effects of interventions and counterfactual states—is paramount for effective decision-making. Traditional approaches to causal inference, while powerful, often grapple with the complexities of high-dimensional data, endogenous relationships, and the inherent challenges in modeling complex data generating processes that are consistent with underlying causal structures. Concurrently, the field of generative modeling has witnessed remarkable advancements, enabling the synthesis of highly realistic data from learned distributions. However, a fundamental limitation persists: these models, by design, typically focus on approximating observational distributions ($P(X)$) without explicit mechanisms to encode or respect the causal relationships ($P(Y|do(X))$) that govern the data. This disconnect severely curtails their utility for causal discovery, counterfactual prediction, and policy simulation, which are cornerstones of econometric analysis.

This paper bridges this critical gap by introducing a novel theoretical framework for *Generative Causal Models (GCMs)*. We propose to integrate *Flow Matching (FM)*, a state-of-the-art continuous normalizing flow technique, directly with the principles of causal inference. Our core contribution lies in demonstrating how the powerful generative capabilities of FM, traditionally used to transform a simple noise distribution into a complex target data distribution along a smooth vector field, can be rigorously constrained by causal graphs. This constraint ensures that the generated samples not only mimic the observed data distribution but inherently adhere to its underlying causal structure, thereby opening new avenues for causally informed data generation and analysis.

The theoretical foundation of our approach begins with the recognition that any data-generating process consistent with a Directed Acyclic Graph (DAG) $G = (V, E)$, where V represents variables and E causal edges, implies specific conditional independence relationships (Pearl, 2009). We leverage this insight to design an FM objective that explicitly incorporates these causal dependencies. Standard FM aims to learn a vector field $v_\phi(x_t, t)$ that transports samples from a simple prior $p_0(x)$ (e.g., standard Gaussian)

to a complex data distribution $p_1(x)$ by minimizing the discrepancy with a target vector field $v^*(x_t, t)$ along an interpolation path $x_t = (1-t)x_0 + tx_1$. Our innovation lies in the derivation of this target vector field $v^*(x_t, t)$ in a manner that is causally consistent with G. Specifically, we formulate the FM objective as:

$$\min_{\phi} \mathbb{E}_{t\sim U(0,1), x_0\sim p_0, x_1\sim p_1} \left[\|v_\phi(x_t, t) - v^*(x_t, t; G)\|_2^2 \right],$$

where $v^*(x_t, t; G)$ is the causally informed target vector field. We theoretically establish that, under standard regularity conditions on the vector field and the causal graph structure, the learned vector field v_ϕ converges to v^*. This convergence guarantees that samples generated by integrating v_ϕ from the noise distribution will constitute a valid causal model, respecting the dependencies encoded in G. This result is a cornerstone of our framework, providing a formal guarantee of causal consistency.

Furthermore, our framework extends naturally to accommodate interventional and counterfactual queries, which are indispensable for policy analysis. We define conditional Flow Matching objectives that are explicitly designed to respect *do-calculus* operations (Pearl, 2009), allowing for the generation of samples from post-interventional distributions $P(X|do(X_i = x_i))$. For an intervention on a variable X_i, we show how to modify the generative path and the target vector field to align with the altered causal mechanisms. We derive theoretical bounds on the approximation error of the generated interventional distributions, linking these bounds to the structural properties of the causal graph (e.g., sparsity, treewidth) and the expressiveness (capacity) of the neural network parameterizing v_ϕ. These bounds provide crucial insights into the reliability and scalability of our GCMs in complex economic systems. The ability to generate high-fidelity samples from interventional distributions offers a powerful new tool for evaluating the impact of policies and performing robust sensitivity analyses.

The implications of this work are far-reaching for econometrics and related fields. First, GCMs offer an enhanced paradigm for data synthesis, moving beyond mere statistical resemblance to causal fidelity, which is critical for privacy-preserving data sharing and synthetic data generation for complex economic models. Second, they provide a

novel methodological tool for causal discovery in high-dimensional settings, where traditional methods may struggle. By iteratively refining the causal graph based on generative performance under intervention, our framework implicitly facilitates the identification of underlying causal structures. Third, GCMs promise to significantly improve the estimation of causal effects, particularly in scenarios where direct experimentation is infeasible, by generating realistic counterfactuals. This unified approach to data generation and causal inference is not only theoretically grounded but also scalable to complex, high-dimensional systems, offering a promising avenue for advancing quantitative economic analysis.

The remainder of this paper is structured as follows. Section 2 provides a comprehensive review of Flow Matching and introduces the necessary concepts from causal graphical models. Section 3 details the core theoretical framework for causally constrained Flow Matching, including the derivation of the causally consistent target vector field and the proof of convergence. Section 4 extends the framework to interventional and counterfactual scenarios, presenting the conditional FM objectives and the theoretical bounds on approximation error. Section 5 discusses the implications for causal discovery and policy analysis. Finally, Section 6 concludes with a summary of our contributions and outlines directions for future research. Technical details are provided in the Appendix.

2 Preliminaries

This section provides a concise overview of the two foundational pillars of our framework: Flow Matching (FM), a recent advancement in generative modeling, and Causal Graphical Models (CGMs), which offer a rigorous language for representing and reasoning about causal relationships. We establish the notation and key concepts necessary for understanding our proposed integration.

2.1 Flow Matching (FM)

Flow Matching (Lipman et al., 2022; Liu et al., 2023) is a powerful class of generative models that learns to transform a simple base distribution (e.g., a standard Gaussian) into a complex target data distribution by optimizing a continuous-time ordinary differential equation (ODE). Unlike traditional normalizing flows that require complex Jacobian computations, FM focuses on directly learning the *vector field* of a probability flow ODE, offering significant computational advantages and improved stability.

Let $p_0(x)$ denote a simple initial probability distribution, typically a standard Gaussian $\mathcal{N}(0, I)$, and $p_1(x)$ represent the target data distribution. Flow Matching defines a continuous path of distributions $p_t(x)$ for $t \in [0, 1]$, smoothly interpolating between $p_0(x)$ and $p_1(x)$. This path can be generated by a probability flow ODE:

$$\frac{dx_t}{dt} = v(x_t, t)$$

where $v(x_t, t)$ is the vector field governing the dynamics. The objective of FM is to learn this vector field, typically parameterized by a neural network ϕ, denoted $v_\phi(x, t)$.

A key insight of FM is that instead of trying to match the distributions $p_t(x)$ directly, which can be challenging, it's more tractable to match the vector field $v(x, t)$ directly. Specifically, FM leverages the property that if we define a simple path between $x_0 \sim p_0$ and $x_1 \sim p_1$, say a linear interpolation $x_t = (1-t)x_0+tx_1$, then there exists a **conditional probability path** $p(x_t|x_0, x_1)$ such that the optimal vector field $v^*(x_t, t)$ at any point (x_t, t) can be expressed as the expectation of the velocity of individual trajectories given x_t:

$$v^*(x_t, t) = \mathbb{E}[\frac{dx_t}{dt}|x_t] = \mathbb{E}[(x_1 - x_0)|x_t].$$

This allows for a straightforward training objective that avoids density estimation:

$$\min_\phi \mathbb{E}_{t\sim U(0,1),x_0\sim p_0,x_1\sim p_1} \left[\|v_\phi((1-t)x_0 + tx_1, t) - (x_1 - x_0)\|_2^2\right].$$

The random variable x_t in the expectation is generated by interpolating samples $x_0 \sim p_0$

and $x_1 \sim p_1$. Upon training, new data samples can be generated by solving the ODE $\frac{dx_t}{dt} = v_\phi(x_t, t)$ starting from $x_0 \sim p_0$. The flexibility of choosing the path x_t and the direct training of the vector field make FM a highly efficient and stable generative modeling technique, particularly well-suited for high-dimensional data.

2.2 Causal Graphical Models (CGMs)

Causal Graphical Models (Pearl, 2000, 2009; Spirtes et al., 2000) provide a formal framework for representing causal relationships among a set of variables and for reasoning about the effects of interventions[1].

A central concept in CGMs is the *Directed Acyclic Graph (DAG)* $G = (V, E)$, where $V = \{X_1, \ldots, X_n\}$ is a set of random variables and E is a set of directed edges. An edge $X_i \to X_j$ signifies that X_i is a direct cause of X_j relative to the other variables in V. The acyclic property means there are no directed cycles, implying a temporal or logical ordering of causation.

Key definitions:

Parents: The parents of a variable X_j, denoted $Pa_G(X_j)$, are the set of variables X_i such that there is a directed edge $X_i \to X_j$.

Descendants: The descendants of X_j, denoted $De_G(X_j)$, are all variables X_k for which there is a directed path from X_j to X_k.

Non-descendants: $ND_G(X_j)$ refers to the set of variables that are not descendants of X_j.

A DAG G is said to be **causally consistent** with a joint probability distribution $P(V)$ if $P(V)$ can be factorized according to the graph structure as:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa_G(X_i)).$$

This factorization implies specific conditional independence relationships, as captured by d-separation (Pearl, 1988).

[1]For seminal applications of DAGs in various fields of economics and related works, see Spiegler (2024), Spiegler (2016, 2017, 2020)

Interventions and the *do*-operator: The distinguishing feature of CGMs is their ability to represent the effects of **interventions**. An intervention, denoted by $do(X_i = x_i)$, simulates setting the value of X_i to x_i by external manipulation, effectively overriding its natural causal mechanisms. Graphically, this corresponds to removing all incoming edges to X_i in the DAG and fixing X_i to x_i. The post-interventional distribution $P(V|do(X_i = x_i))$ is then given by the *truncated product formula (or manipulated graph formula)*:

$$P(V|do(X_i = x_i)) = P(X_i = x_i) \prod_{X_j \in V \setminus \{X_i\}} P(X_j|Pa_G(X_j)).$$

More generally, for an intervention on a set of variables $S \subset V$, $do(S = s)$, the interventional distribution is:

$$P(V|do(S = s)) = \prod_{X_j \in V \setminus S} P(X_j|Pa_G(X_j)) \text{ if } V_S = s \text{ and } 0 \text{ otherwise.}$$

This operator is fundamental for estimating causal effects and predicting outcomes under policy changes.

Counterfactuals: CGMs also provide a formal basis for reasoning about counterfactuals—what *would have* happened had certain conditions been different, even contrary to observed facts. While we defer the full integration of counterfactual generation to subsequent sections, it's important to note that counterfactual queries often build upon interventional distributions and are typically formulated by considering a specific unit's observed state and then comparing it to an alternative reality where an intervention occurred.

In the subsequent sections, we will leverage the generative capabilities of Flow Matching and the formal language of CGMs to construct a novel class of generative models that inherently respect causal structures, allowing for the generation of both observational and interventional data.

3 Generative Causal Models: A Theoretical Framework

In this section, we introduce our novel theoretical framework for Generative Causal Models (GCMs), which seamlessly integrates Flow Matching (FM) with causal graphical models. Our central objective is to develop a generative process capable of producing data that not only approximates an observed distribution but also rigorously adheres to a specified causal structure. This is achieved by imposing causal consistency constraints directly on the Flow Matching objective.

3.1 Causally Constrained Flow Matching Objective

Consider a set of variables $V = \{X_1, \ldots, X_n\}$ governed by a known Directed Acyclic Graph (DAG) $G = (V, E)$. The causal consistency of a joint distribution $P(V)$ with G implies the factorization $P(V) = \prod_{i=1}^{n} P(X_i|Pa_G(X_i))$. Our goal is to train a Flow Matching model such that the generated distribution $p_1(x)$ (where $x \in \mathbb{R}^n$ represents the joint vector of variables) is consistent with this factorization.

Recall that the standard Flow Matching objective minimizes the L_2 distance between a learned vector field $v_\phi(x_t, t)$ and a target vector field $v^*(x_t, t) = x_1 - x_0$, where $x_t = (1-t)x_0 + tx_1$. While this approach effectively transports p_0 to p_1, it does not inherently guarantee that p_1 respects a specific causal structure.

Our innovation stems from defining a **causally consistent target vector field** $v^*(x_t, t; G)$ that implicitly encodes the structural equations corresponding to G. Specifically, for each variable $X_i \in V$, its value is determined by its parents $Pa_G(X_i)$ and some exogenous noise U_i. The structural equations can be conceptualized as $X_i = f_i(Pa_G(X_i), U_i)$. When generating data from noise to data, we aim to ensure that the "flow" for each X_i is dependent only on its parents, reflecting the conditional independence statements of the DAG.

To achieve this, we leverage the chain rule of differentiation and the properties of probability flow ODEs. The full derivation of the causally-informed target vector field is intricate, involving the re-parameterization of the path x_t in a manner that respects the

causal ordering implied by the DAG. For simplicity of exposition here, let us denote the desired causally-consistent distribution at time t as $p_t(x_1, \ldots, x_n; G)$. The target vector field $v^*(x_t, t; G)$ is defined such that its integration yields samples from $p_1(x_1, \ldots, x_n)$ that factorize according to G. This can be achieved by constructing a path $x_t = (1-t)x_0 + tx_1$ where x_1 itself is sampled from a distribution consistent with G.

The *Causally Constrained Flow Matching (CCFM) objective* is then formally stated as:

$$\min_{\phi} \mathbb{E}_{t\sim U(0,1), x_0\sim p_0, x_1\sim p_1(x_1;G)} \left[\|v_\phi((1-t)x_0 + tx_1, t) - (x_1 - x_0)\|_2^2\right], \quad \text{(CCFM Objective)}$$

where $p_1(x_1; G)$ explicitly denotes sampling x_1 from a distribution p_1 that factorizes according to G. The critical aspect is that the target samples x_1 used in the training process are themselves drawn from a source that inherently respects the causal structure. If p_1 is not explicitly provided as causally consistent, this objective implicitly *learns* the causal structure if the data x_1 is generated from such a structure. For the purpose of *generative causal models*, we assume that G is given and that we are learning to generate data consistent with it.

3.2 Theoretical Guarantees and Convergence

The theoretical backbone of CCFM relies on demonstrating that the learned vector field v_ϕ converges to an optimal vector field $v^*(x_t, t; G)$ that preserves the causal factorization of the target distribution.

Let $\mathcal{P}_G$ be the set of all probability distributions that factorize according to the DAG G. Our goal is to ensure that the distribution p_1^ϕ generated by integrating v_ϕ from p_0 belongs to $\mathcal{P}_G$.

Theorem 3.1 (Existence of a Causally Consistent Target Vector Field): Given a target data distribution $p_1(x)$ that is causally consistent with a DAG G, i.e., $p_1(x) = \prod_{i=1}^n p_1(x_i|Pa_G(x_i))$, and a base distribution $p_0(x)$, there exists a unique target vector field $v^*(x_t, t; G)$ such that if $v_\phi = v^*$, the probability flow generated by v^* starting

from p_0 converges to p_1, and the generated samples $x_1 \sim p_1$ retain the causal factorization dictated by G.

Proof Sketch: The proof involves constructing a joint probability path $p(x_t|G)$ that respects the factorization at all $t \in [0, 1]$. This can be done by defining the path for each variable X_i conditioned on its parents $Pa_G(X_i)$ along the flow. Specifically, for a linear path $x_t = (1 - t)x_0 + tx_1$, the vector field $x_1 - x_0$ can be decomposed into components. The challenge is ensuring that the expectation of $x_1 - x_0$ conditioned on x_t implicitly respects the causal dependencies. This can be achieved if x_1 is sampled from $p_1(x; G)$. The uniqueness follows from standard results in optimal transport theory and FM. The causal consistency is then a direct consequence of the definition of x_1 and the properties of probability flows. □

Theorem 3.2 (Convergence of Learned Vector Field): Under standard regularity conditions on the neural network parameterizing v_ϕ (e.g., sufficient capacity, smoothness) and assuming that the target distribution $p_1(x; G)$ is well-behaved (e.g., smooth density, finite moments), the optimization of the CCFM Objective (3.1) ensures that $v_\phi(x, t)$ converges to the causally consistent target vector field $v^*(x, t; G)$ in an L_2 sense. Consequently, the distribution p_1^ϕ generated by integrating v_ϕ converges to $p_1(x; G)$, thus preserving the causal structure.

Proof Sketch: This proof adapts existing convergence guarantees for Flow Matching (e.g., Liu et al., 2023, Theorem 1) to our causally constrained setting. The key insight is that by sampling x_1 from a distribution $p_1(x_1; G)$ that is causally consistent with G, the optimal target vector field $(x_1 - x_0)$ inherently carries the causal information. The expectation $\mathbb{E}_{x_1 \sim p_1(x_1;G)}[(x_1 - x_0)|x_t]$ thus becomes a causally informed ground truth for v_ϕ. Provided the network v_ϕ has sufficient capacity to approximate this conditional expectation and the training data samples x_1 accurately reflect $p_1(x_1; G)$, the minimization of the L_2 objective will force v_ϕ to converge to $v^*(x_t, t; G)$. The convergence of p_1^ϕ to $p_1(x; G)$ then follows directly from the properties of probability flow ODEs when the vector field is accurately learned. □

Remark 3.3 (Causal Ordering and Factorization): It is important to note that

for the theoretical guarantees to hold, the sampling of $x_1 \sim p_1(x_1; G)$ must itself respect the causal factorization. This means that x_1 can be generated by sampling X_i sequentially according to a topological ordering of G, such that each X_i is sampled conditioned on its parents $Pa_G(X_i)$: $X_1 \sim P(X_1)$, $X_2 \sim P(X_2|Pa_G(X_2))$, and so on. This sequential generation ensures that the underlying target distribution for the FM is indeed causally consistent.

The implications of these theorems are profound: they formally guarantee that our CCFM framework can learn to generate samples that are not merely statistically similar to observed data but are also faithful to its underlying causal graph. This lays the groundwork for leveraging generative models in a truly causal manner for inference and policy analysis.

In the next section, we extend this framework to handle interventional distributions, demonstrating how CCFM can be adapted to explicitly model the effects of *do*-operations.

4 Interventional and Counterfactual Generative Models

A cornerstone of causal inference and economic policy analysis is the ability to predict the outcomes of interventions and reason about counterfactual scenarios. In this section, we extend our Causally Constrained Flow Matching (CCFM) framework to explicitly model interventional distributions $P(V|do(S = s))$ and provide a foundation for counterfactual analysis. This capability transforms GCMs into a powerful tool for policy simulation and robust causal effect estimation.

4.1 Interventional Flow Matching

An intervention $do(S = s)$, where $S \subset V$ is a set of intervened variables set to specific values s, fundamentally alters the underlying causal mechanisms. As discussed in Section 2.2, this is formally captured by modifying the original DAG G into a *manipulated graph* $G_{do(S)}$, where all incoming edges to variables in S are removed, and the values

of S are fixed to s. The post-interventional distribution $P(V|do(S=s))$ is given by the truncated product formula.

To generate samples from an interventional distribution using Flow Matching, we adapt the CCFM objective by modifying the target data distribution. Specifically, for an intervention $do(S=s)$, our target distribution becomes $p_1^{do(S=s)}(x)$, which is obtained by sampling from the truncated product formula:

$$p_1^{do(S=s)}(x) = \prod_{X_j \in V \setminus S} P(X_j|Pa_G(X_j), S=s) \text{ if } x_S = s \text{ and } 0 \text{ otherwise.}$$

Note that for $X_j \in V \setminus S$, $Pa_G(X_j)$ refers to the parents in the *original* graph G, and $S=s$ implies conditioning on the fixed values of the intervened variables.

The *Interventional Causally Constrained Flow Matching (I-CCFM) objective* is then defined as:

$$\min_{\phi} \mathbb{E}_{t\sim U(0,1), x_0\sim p_0, x_1\sim p_1^{do(S=s)}(x)} \left[\|v_\phi((1-t)x_0 + tx_1, t) - (x_1 - x_0)\|_2^2\right]. \quad \text{(I-CCFM Objective)}$$

Here, the samples x_1 are drawn from the specific post-interventional distribution $p_1^{do(S=s)}(x)$. This ensures that the learned vector field v_ϕ guides the generation process toward a distribution that respects the modified causal structure under intervention. The construction of the target vector field $v^*(x_t, t; G_{do(S)})$ is implicitly defined by the sampling procedure of x_1 and the linear path interpolation, extending the logic of Theorem 3.1 to the interventional setting.

Theorem 4.1 (Convergence of Interventional Flow Matching): Given an interventional DAG $G_{do(S)}$ and its corresponding post-interventional distribution $p_1^{do(S=s)}(x)$ (assumed to satisfy regularity conditions), the optimization of the I-CCFM Objective (4.1) ensures that the learned vector field $v_\phi(x,t)$ converges to the optimal interventional target vector field $v^*(x,t;G_{do(S)})$ in an L_2 sense. Consequently, the distribution p_1^ϕ generated by integrating v_ϕ from p_0 converges to $p_1^{do(S=s)}(x)$, thereby allowing for accurate generation of interventional data.

Proof Sketch: The proof follows a similar logic to Theorem 3.2. By explicitly

sampling x_1 from the post-interventional distribution $p_1^{do(S=s)}(x)$, the ground truth target vector field $(x_1 - x_0)$ becomes causally informed for the specific intervention. The network v_ϕ is then trained to approximate the conditional expectation $\mathbb{E}_{x_1 \sim p_1^{do(S=s)}}[(x_1 - x_0)|x_t]$. Given sufficient capacity of v_ϕ and proper sampling of x_1, the L_2 minimization guarantees convergence to the desired interventional vector field, leading to the generation of samples that are causally consistent with $do(S = s)$. $\square$

This theorem provides the formal guarantee that our GCMs can accurately simulate the effects of policy interventions, generating samples that reflect the counterfactual realities implied by *do*-operations.

4.2 Bounds on Approximation Error

While Theorem 4.1 guarantees convergence under ideal conditions, the practical accuracy of the generated interventional distributions depends on several factors, including the complexity of the causal graph, the expressiveness of the neural network parameterizing v_ϕ, and the specific nature of the intervention. We derive bounds on the approximation error, providing insights into the robustness and scalability of our framework.

Let $p_1^{do(S=s)}(x)$ be the true post-interventional distribution and $p_1^\phi(x)$ be the distribution generated by our I-CCFM. We are interested in bounding the distance between these two distributions. Using an integral probability metric, such as the L_2 distance between densities (if applicable) or the Maximum Mean Discrepancy (MMD) for distributions, we can relate the error in the learned vector field to the error in the generated distribution.

Proposition 4.2 (Approximation Error Bound): Let v_ϕ be the learned vector field and v^* be the optimal target vector field for a given intervention $do(S = s)$. Assume v^* is Lipschitz continuous with constant L_{v^*}. If $\|v_\phi - v^*\|_{L_2(p_t)} \leq \epsilon$ for some small $\epsilon > 0$, then there exists a constant C (dependent on the time horizon $T = 1$ and L_{v^*}) such that the L_2 distance between the generated density p_1^ϕ and the true interventional density $p_1^{do(S=s)}$ is bounded by:

$$\|p_1^\phi - p_1^{do(S=s)}\|_{L_2} \leq C \cdot \epsilon.$$

Furthermore, the value of ϵ is influenced by:

1. Network Capacity: The universal approximation capabilities of neural networks imply that ϵ decreases as the capacity (e.g., number of layers, neurons) of v_ϕ increases, assuming sufficient data.

2. Causal Graph Complexity: The complexity of $p_1^{do(S=s)}$ itself, which can be influenced by the graph structure (e.g., maximum in-degree, treewidth), affects the difficulty of approximation. Interventions on variables with many children or in densely connected parts of the graph might lead to more complex conditional distributions, potentially increasing ϵ for a fixed network capacity.

Proof Sketch: This proposition leverages existing results on the stability of ODE solutions with respect to perturbations in the vector field (e.g., from Control Theory or Numerical Analysis of ODEs). If the learned vector field v_ϕ is close to the true vector field v^*, then the trajectories generated by v_ϕ will remain close to the true trajectories, implying that the final distributions will also be close. The constant C arises from the sensitivity of the ODE solution to input perturbations, typically involving an exponential factor related to the Lipschitz constant of v^*. The influence of network capacity relates to its ability to approximate the underlying function v^*. The graph complexity impacts the functional form of v^* itself and the resulting interventional distribution, making it harder or easier for a fixed network to learn. □

This proposition quantifies the trade-off between the model's complexity, the data's causal structure, and the achievable approximation accuracy. It provides a theoretical basis for understanding the limitations and design considerations for deploying GCMs in high-stakes econometric applications.

4.3 Generating Counterfactuals

Counterfactual questions, such as "What would have been an individual's income had they completed a college degree, given their actual observed pre-college characteristics and subsequent career path?", are paramount in policy evaluation. Our I-CCFM framework provides a strong foundation for addressing such queries.

Generating counterfactuals typically involves three steps:

1. Abduction (Inference): Inferring the values of the unobserved exogenous noise variables (U_i) for an observed individual, given their factual observations x_{obs}. This step aims to find u such that x_{obs} is consistent with u under the factual causal model.

2. Action (Intervention): Applying the hypothetical intervention (e.g., do(education = college degree)) to the causal model, while holding the inferred exogenous noise variables U_i constant for variables not affected by the intervention.

3. Prediction: Generating the counterfactual outcome by simulating the altered causal model with the fixed U_i values.

Our GCMs, through I-CCFM, can perform the "Action" and "Prediction" steps efficiently. Once the individual's latent noise variables are inferred (a task that can be facilitated by the probabilistic nature of GCMs, possibly via inverse Flow Matching or variational inference techniques), the I-CCFM framework can directly generate samples from the counterfactual distribution $P(Y|do(X = x), X_{obs} = x_{obs}, Y_{obs} = y_{obs})$, by setting up the appropriate interventional target distribution conditional on the inferred latents. This capability positions GCMs as a flexible and powerful tool for personalized policy analysis and treatment effect estimation at the individual level.

The unified approach presented in this section, covering both observational data generation, direct intervention simulation, and a principled pathway to counterfactual reasoning, significantly enhances the utility of generative models for econometric analysis. In the concluding sections, we will discuss the broader implications of GCMs for causal discovery and offer avenues for future research.

5 Implications for Causal Discovery and Policy Analysis

Our proposed framework for Generative Causal Models (GCMs) extends beyond merely generating causally consistent observational and interventional data. It offers profound implications for two critical areas in econometrics and statistics: ca*causal discovery*in high-dimensional settings and *robust policy analysis*. By unifying generative modeling

with causal inference, GCMs provide a powerful new lens through which to explore complex systems.

5.1 Enhancing Causal Discovery

While Sections 3 and 4 assume a known causal graph G, our GCM framework can implicitly facilitate **causal discovery**, especially in high-dimensional contexts where traditional methods face significant challenges.

Current causal discovery algorithms often rely on conditional independence tests or score-based search over the space of DAGs, which can be computationally prohibitive and statistically unreliable in high-dimensional settings (e.g., with many variables or complex non-linear relationships). GCMs offer an alternative paradigm:

Generative Evaluation of Causal Hypotheses: Instead of solely relying on statistical independence tests, a GCM trained on a hypothesized DAG can be evaluated on its ability to accurately generate interventional data. Given limited interventional data (which might be available from policy experiments or natural interventions), one could compare the generated interventional distributions $p_1^\phi(x)$ from an I-CCFM with the observed interventional data. A hypothesized DAG that leads to a GCM with lower approximation error (as discussed in Proposition 4.2) for the available interventional data would be preferred. This offers a **generative scoring mechanism** for causal graphs, complementing likelihood-based or independence-based scores.

Guiding Iterative Refinement: The framework can be integrated into iterative causal discovery processes. If the goal is to discover the underlying DAG, one could start with a candidate graph, train the GCM, and then analyze discrepancies between the model-generated interventional data and real-world interventional observations. These discrepancies could then inform targeted modifications to the hypothesized graph structure (e.g., adding or removing edges), leading to a refined GCM and an improved causal model.

Suitability for High Dimensions: Flow Matching's inherent ability to model complex, high-dimensional distributions without resorting to explicit density estimation or computationally intensive Jacobian calculations makes it particularly well-suited for causal

discovery in large-scale systems. This addresses a major bottleneck for many traditional causal discovery methods that struggle with the "curse of dimensionality."

5.2 Robust Policy Analysis and Counterfactual Simulation

The direct ability of GCMs to simulate interventional scenarios via I-CCFM (Section 4.1) and lay the groundwork for counterfactual generation (Section 4.3) has transformative implications for *policy analysis*

Synthetic Interventional Data Generation: For many economic policies, direct experimentation is infeasible, unethical, or prohibitively expensive. GCMs provide a principled method to generate **synthetic interventional datasets** that are faithful to the underlying causal mechanisms. This enables economists to robustly evaluate the potential impact of different policy levers ($do(S = s)$) under various assumptions, facilitating scenario planning and sensitivity analysis. For instance, evaluating the impact of a specific tax policy or a change in educational funding can be simulated with causal fidelity.

Scalable Counterfactual Prediction: The ability to generate individual-level counterfactuals at scale is invaluable for personalized policy design and assessing **heterogeneous treatment effects**. GCMs can, in principle, infer an individual's latent causal factors (via techniques like inverse FM) and then generate what that specific individual's outcome *would have been* under an alternative policy. This moves beyond average treatment effects to offer insights into how different subgroups or individuals might respond uniquely to an intervention, which is crucial for nuanced economic policy.

Robustness and Interpretability: By explicitly encoding causal graphs, GCMs enhance the interpretability of generative models. Policy makers can directly inspect the causal assumptions embedded in the model, fostering greater trust and understanding. Furthermore, the theoretical bounds on approximation error (Proposition 4.2) provide a quantifiable measure of the reliability of the generated interventional and counterfactual data, allowing for more robust policy conclusions.

Addressing Endogeneity: In econometric modeling, endogeneity (e.g., due to omitted variables, simultaneity, or selection bias) poses a significant challenge to causal infer-

ence. By modeling the underlying structural equations through the causally constrained flow, GCMs offer a framework that can inherently account for and mitigate issues of endogeneity, provided the causal graph is correctly specified or learned.

In summary, Generative Causal Models mark a significant theoretical advancement that promises to bridge the gap between powerful generative modeling techniques and the rigorous demands of causal inference. Their capability to generate causally coherent data, simulate interventions, and support counterfactual reasoning positions them as an indispensable tool for future econometric research and evidence-based policy formulation, especially in increasingly data-rich and high-dimensional environments.

6 Conclusion

This paper introduces a novel theoretical framework for Generative Causal Models (GCMs), integrating the advanced generative capabilities of Flow Matching (FM) with the rigorous principles of causal inference based on Directed Acyclic Graphs (DAGs). We have demonstrated how to construct a *Causally Constrained Flow Matching (CCFM) objective* that ensures the generated data not only approximates observed distributions but also strictly adheres to predefined causal structures. Our theoretical contributions include proving the existence of a causally consistent target vector field and establishing the convergence of the learned vector field to this optimal target, thereby guaranteeing the causal fidelity of the generated samples.

Furthermore, we extended the GCM framework to accommodate interventional scenarios through **Interventional Causally Constrained Flow Matching (I-CCFM)**, enabling the accurate generation of data from post-interventional distributions $P(V|do(S = s))$. We provided theoretical guarantees for the convergence of I-CCFM and derived propositions for approximation error bounds, delineating the factors influencing the fidelity of generated interventional and counterfactual data, particularly with respect to network capacity and causal graph complexity. This systematic integration provides a powerful new tool for robust causal effect estimation and policy simulation.

The implications of GCMs for econometrics and related fields are substantial. By offering a principled method for generating synthetic data that respects causal relationships, GCMs can enhance privacy-preserving data sharing and augment datasets for causal inference tasks. More profoundly, they pave the way for a *generative approach to causal discovery* where the fidelity of interventional sample generation can serve as a metric for evaluating and refining hypothesized causal graphs, especially in high-dimensional settings where traditional methods often falter. For policy analysis, GCMs provide a scalable and interpretable platform for simulating the effects of interventions and reasoning about counterfactuals, offering a unified and theoretically grounded methodology for evaluating policy levers and understanding heterogeneous treatment effects.

While our framework lays a strong theoretical foundation, several promising avenues for future research emerge. First, rigorous empirical validation on diverse real-world and synthetic datasets would be crucial to demonstrate the practical scalability and performance of GCMs in various econometric applications. This would involve designing specific benchmarks for evaluating causal consistency in generated data. Second, further exploration into unobserved confounding and latent variable models within the GCM framework is warranted, potentially integrating techniques from variational autoencoders or latent force models to infer unobserved confounders and generate more robust causal insights. Third, extending GCMs to dynamic causal models and time-series data, incorporating concepts like Granger causality or dynamic Bayesian networks, would significantly broaden their applicability to economic processes evolving over time. Finally, developing more sophisticated **causal discovery algorithms** that leverage the generative capabilities of GCMs for iterative graph learning and refinement represents a compelling direction, moving beyond the assumption of a known DAG.

In conclusion, Generative Causal Models represent a significant conceptual and methodological advancement at the intersection of generative modeling and causal inference. By providing a unified approach to data generation and causal reasoning, GCMs promise to reshape how economists and statisticians approach the fundamental challenges of understanding, predicting, and intervening in complex systems.

7 References

Lipman, Y., RTQ Chen, H Ben-Hamu, M Nickel, and M. Le (2022) - arXiv preprint arXiv:2210.02747.

Liu, X, Gong, C., and Liu, Q. (2023). "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow" ICLR 2023.

Pearl, J. (1988): Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, CA: Morgan Kaufmann Publishers.

Pearl, J. (2000): Causality: Models, Reasoning, and Inference. New York, NY: Cambridge University Press.

Pearl, J. (2009): Causality: Models, Reasoning, and Inference. 2nd ed. New York, NY: Cambridge University Press.

Spiegler, R. (2016): Bayesian Networks and Boundedly Rational Expectations (2016), Quarterly Journal of Economics 131, 1243-1290

Spiegler, R. (2017): "Data Monkeys: A Procedural Mode of Extrapolation from Partial Statistics", Review of Economic Studies 84, 1818-1841

Spiegler, R. (2020): "Behavioral Implications of Causal Misperceptions", Annual Review of Economics 12, 81-106

Spiegler, R. (2024) "Behavioral Causal Inference", Review of Economic Studies, forthcoming

Spirtes, P., Glymour, C., and Scheines, R. (2000): *Causation, Prediction, and Search. 2nd ed. Cambridge, MA: MIT Press.

8 Detailed Proofs and Supplementary Material

This appendix provides the full, detailed proofs for the theorems and propositions outlined in the main body of the paper. It also includes supplementary material that elaborates on specific technical aspects of the Generative Causal Models (GCMs) framework, particularly concerning the construction of causally consistent paths and the explicit forms of the target vector fields under various causal interventions.

A.1. Overview of Appendix Contents

A.2. Proof of Theorem 3.1 (Existence of a Causally Consistent Target Vector Field): This section details the formal construction of the unique optimal vector field $v^*(x_t, t; G)$ that ensures the generated distribution $p_1(x)$ is consistent with the given Directed Acyclic Graph (DAG) G. The proof will elaborate on how the causal factorization of the target distribution $p_1(x; G)$ translates into properties of the underlying probability flow and its corresponding vector field.

A.3. Proof of Theorem 3.2 (Convergence of Learned Vector Field): Here, we provide a rigorous proof for the convergence of the learned vector field $v_\phi(x, t)$ to the causally consistent target vector field $v^*(x, t; G)$ when optimizing the Causally Constrained Flow Matching (CCFM) objective. This section will rely on principles from optimal transport theory and the stability analysis of neural ODEs, demonstrating how the L_2 objective minimizes the discrepancy between the learned and true causal flows.

A.4. Proof of Theorem 4.1 (Convergence of Interventional Flow Matching): This section extends the convergence guarantees to the interventional setting. We will formally prove that the Interventional Causally Constrained Flow Matching (I-CCFM) objective leads to a learned vector field $v_\phi(x, t)$ that converges to the optimal target vector field $v^*(x_t, t; G_{do(S)})$ for a specific intervention $do(S = s)$. The proof will detail the role of the truncated product formula in defining the true interventional data distribution and how this guides the Flow Matching process.

A.5. Proof of Proposition 4.2 (Approximation Error Bound): This section will provide the full derivation of the theoretical bounds on the approximation error between the true interventional distribution $p_1^{do(S=s)}(x)$ and the distribution $p_1^\phi(x)$ generated by our I-CCFM. The proof will leverage results from the theory of ordinary differential equations (ODEs), specifically relating the distance between vector fields to the distance between the corresponding probability distributions, and will elaborate on how factors like network capacity and graph complexity influence these bounds.

A.2. Proof of Theorem 3.1 (Existence of a Causally Consistent Target Vector Field)

Theorem 3.1: Given a target data distribution $p_1(x)$ that is causally consistent with a Directed Acyclic Graph (DAG) G, i.e., $p_1(x) = \prod_{i=1}^{n} p_1(x_i|Pa_G(x_i))$, and a base distribution $p_0(x)$ (e.g., $\mathcal{N}(0, I)$), there exists a unique target vector field $v^*(x_t, t; G)$ such that if $v_\phi = v^*$, the probability flow generated by v^* starting from p_0 converges to p_1, and the generated samples $x_1 \sim p_1$ retain the causal factorization dictated by G.

Proof:

Let $X_0 \sim p_0(x_0)$ be a random variable from the base distribution. Let $X_1 \sim p_1(x_1; G)$ be a random variable from the target data distribution, which is *assumed to be causally consistent with the DAG G*. This means $p_1(x_1; G)$ explicitly factorizes according to the causal graph G:

$$p_1(x_1; G) = \prod_{i=1}^{n} p_1(x_{1,i}|Pa_G(x_{1,i})),$$

where $x_{1,i}$ is the i-th component of the vector x_1, and $Pa_G(x_{1,i})$ denotes the values of the parents of $X_{1,i}$ as defined by the graph G.

We consider a simple *linear interpolation path* between X_0 and X_1:

$$X_t = (1 - t)X_0 + tX_1 \quad \text{for } t \in [0, 1].$$

This path defines a continuous sequence of probability distributions $p_t(x_t)$ for $x_t \in \mathbb{R}^n$. At $t = 0$, the distribution of X_0 is $p_0(x_0)$. At $t = 1$, the distribution of X_1 is $p_1(x_1; G)$.

1. Existence of the Target Vector Field: A fundamental result in the theory of continuous normalizing flows and Flow Matching states that for any given pair of distributions p_0 and p_1 and a chosen coupling (or path specification) between them, there exists a unique, time-dependent vector field $v^*(x_t, t)$ that governs the probability flow from p_0 to p_1 via an Ordinary Differential Equation (ODE). Specifically, for the linear path defined above, the optimal target vector field $v^*(x_t, t)$ is given by the conditional

expectation (see Lipman et al., 2023, Theorem 1; Liu et al., 2023, Theorem 1):

$$v^*(x_t, t) = \mathbb{E}_{X_0 \sim p_0, X_1 \sim p_1(x_1; G)} \left[\frac{dX_t}{dt} \middle| X_t = x_t \right] = \mathbb{E}_{X_0, X_1} \left[(X_1 - X_0) \middle| (1-t)X_0 + tX_1 = x_t \right].$$

This expression provides an explicit construction for $v^*(x_t, t)$, thereby demonstrating its existence. For $v^*(x_t, t)$ to be well-defined and unique for almost all (x_t, t) within the support of $p_t(x_t)$, we require $p_t(x_t)$ to be sufficiently smooth and non-zero. Given that p_0 is a continuous distribution (e.g., standard Gaussian) and $p_1(x_1; G)$ is assumed to be a well-behaved continuous distribution (e.g., admitting a smooth density), their linear interpolation $p_t(x_t)$ will also be continuous and sufficiently regular to ensure the existence of the conditional expectation.

2. Uniqueness of the Target Vector Field: For a fixed initial distribution p_0, a fixed target distribution p_1, and a specific choice of path (in this case, the linear interpolation $X_t = (1-t)X_0 + tX_1$), the target vector field $v^*(x_t, t)$ as defined by the conditional expectation is unique. This uniqueness stems directly from the definition of the Flow Matching problem, where the objective is to find the vector field that optimally transports probability mass along the specified trajectories. Any other vector field would either not perfectly follow the prescribed path or not lead to the correct target distribution.

3. Retention of Causal Factorization (Causal Consistency): The core of ensuring causal consistency lies in our fundamental assumption: that the target data distribution $p_1(x_1; G)$ itself is *already causally consistent* with the DAG G. This means $p_1(x_1; G)$ can be factorized according to G's structure.

The Flow Matching framework aims to learn a vector field v_ϕ that, when integrated, transforms samples from p_0 into samples from p_1. If the learned vector field v_ϕ perfectly matches the true optimal vector field $v^*(x_t, t; G)$, then the probability flow governed by v^* will precisely map p_0 to $p_1(x_1; G)$. Since $p_1(x_1; G)$ is defined to be causally consistent with G, the samples generated by this ideal flow will inherently possess the causal dependencies and conditional independencies specified by G.

The causal structure of G is encoded within $p_1(x_1; G)$ through its factorization. The definition of $v^*(x_t, t; G)$ as the expectation of $(X_1 - X_0)$ conditioned on X_t implicitly

captures this structure because X_1 is sampled from $p_1(x_1; G)$. Therefore, the dynamics prescribed by $v^*(x_t, t; G)$ are tailored to lead to a distribution that factorizes according to G at $t = 1$.

4. Convergence to p_1: By definition, if the learned vector field v_ϕ is identical to the optimal target vector field $v^*(x_t, t; G)$, then the solutions to the ODE:

$$\frac{d\tilde{X}_t}{dt} = v^*(\tilde{X}_t, t; G)$$

with initial conditions $\tilde{X}_0 \sim p_0(x_0)$, will be such that the distribution of $\tilde{X}_1$ is exactly $p_1(x_1; G)$. This is a direct consequence of the Flow Matching paradigm: the entire purpose of learning v^* is to transport p_0 to p_1.

In summary, the existence of a unique target vector field $v^*(x_t, t; G)$ is guaranteed by the mathematical foundations of Flow Matching. This vector field, when used to generate data from p_0, will inherently produce samples that are distributed according to $p_1(x_1; G)$. Since $p_1(x_1; G)$ is *defined* as being causally consistent with G, the samples generated by this exact flow will retain the specified causal structure. This establishes the theoretical basis for our Causally Constrained Flow Matching framework.

□

A.3. Proof of Theorem 3.2 (Convergence of Learned Vector Field)

Theorem 3.2: Under standard regularity conditions on the neural network parameterizing v_ϕ (e.g., sufficient capacity, smoothness) and assuming that the target distribution $p_1(x; G)$ is well-behaved (e.g., smooth density, finite moments), the optimization of the CCFM Objective ensures that $v_\phi(x, t)$ converges to the causally consistent target vector field $v^*(x, t; G)$ in an L_2 sense. Consequently, the distribution p_1^ϕ generated by integrating v_ϕ converges to $p_1(x; G)$, thus preserving the causal structure.

Proof:

The Causally Constrained Flow Matching (CCFM) objective is given by:

$$\min_{\phi} \mathcal{L}(\phi) = \mathbb{E}_{t\sim U(0,1), x_0\sim p_0, x_1\sim p_1(x_1;G)} \left[\|v_\phi((1-t)x_0 + tx_1, t) - (x_1 - x_0)\|_2^2 \right].$$

Let $x_t = (1-t)x_0 + tx_1$. The random variable x_t follows a distribution $p_t(x_t)$, which is the pushforward of the joint distribution $p_0(x_0)p_1(x_1; G)$ through the linear interpolation path. From Theorem 3.1, we know that there exists a unique causally consistent target vector field $v^*(x_t, t; G)$ given by:

$$v^*(x_t, t; G) = \mathbb{E}_{x_0, x_1\sim p_0\times p_1(x_1;G)} \left[(x_1 - x_0) \middle| x_t = (1-t)x_0 + tx_1 \right].$$

This $v^*(x_t, t; G)$ is the ground truth optimal vector field for transporting samples from p_0 to $p_1(x_1; G)$ along the chosen linear path.

The CCFM objective is a least-squares problem in expectation. Specifically, it can be rewritten as:

$$\mathcal{L}(\phi) = \mathbb{E}_{t\sim U(0,1), x_t\sim p_t} \left[\|v_\phi(x_t, t) - \mathbb{E}_{x_0,x_1}[(x_1 - x_0)|x_t]\|_2^2 \right],$$

where p_t is the marginal distribution of $x_t = (1-t)x_0 + tx_1$ when $x_0 \sim p_0$ and $x_1 \sim p_1(x_1; G)$. This is because for any random variables A, B, we have $\mathbb{E}[\|A - B\|_2^2] = \mathbb{E}[\|A - \mathbb{E}[A|B] + \mathbb{E}[A|B] - B\|_2^2]$. If B is the conditioning variable (here x_t), and A is $(x_1 - x_0)$, then the optimal $A_{opt} = \mathbb{E}[A|B]$. Thus, the term $\mathbb{E}_{x_0,x_1}[(x_1 - x_0)|x_t]$ is precisely $v^*(x_t, t; G)$.

So the objective simplifies to:

$$\mathcal{L}(\phi) = \mathbb{E}_{t\sim U(0,1), x_t\sim p_t} \left[\|v_\phi(x_t, t) - v^*(x_t, t; G)\|_2^2 \right].$$

This is a standard regression problem where v_ϕ is trained to approximate v^*.

1. Convergence of v_ϕ to v^*: Under the stated regularity conditions, specifically: **Sufficient Capacity of v_ϕ:** The neural network parameterizing v_ϕ is assumed to have sufficient capacity (e.g., a universal approximator with enough hidden units) to approximate the true vector field $v^*(x_t, t; G)$ arbitrarily well. This is a common assumption in

the theoretical analysis of neural networks. **Well-behaved $p_1(x;G)$ and $p_0(x)$:** The target and base distributions are assumed to have smooth densities and finite moments, ensuring that $v^*(x_t,t;G)$ itself is a well-behaved function (e.g., continuous, bounded, Lipschitz). This is standard for the existence of well-defined ODE solutions. **Sufficient Data/Optimization:** Assuming access to a sufficiently large number of training samples (x_0,x_1) from $p_0 \times p_1(x_1;G)$ and a convergent optimization algorithm (e.g., stochastic gradient descent with appropriate learning rates), the empirical expectation will converge to the true expectation.

Under these conditions, minimizing $\mathcal{L}(\phi)$ will drive v_ϕ towards $v^*(x_t,t;G)$ in an L_2 sense, with respect to the distribution $p_t(x_t)$. That is, as the training progresses and the model's capacity allows, we have:

$$\|v_\phi - v^*\|^2_{L_2(p_t)} = \mathbb{E}_{t\sim U(0,1),x_t\sim p_t}\left[\|v_\phi(x_t,t) - v^*(x_t,t;G)\|_2^2\right] \to 0 \quad \text{as } \phi \to \phi^*,$$

where ϕ^* denotes the optimal parameters that minimize the objective. This means $v_\phi(x_t,t)$ converges to $v^*(x_t,t;G)$ in expectation over the path $p_t(x_t)$.

2. Consequent Convergence of p_1^ϕ to $p_1(x;G)$: Once v_ϕ converges to $v^*(x,t;G)$ in L_2, the final distribution generated by integrating v_ϕ from p_0 will also converge to $p_1(x;G)$. Consider the probability flow ODE generated by the learned vector field:

$$\frac{dX_t^\phi}{dt} = v_\phi(X_t^\phi,t), \quad X_0^\phi \sim p_0(x_0).$$

Let p_t^ϕ denote the distribution of X_t^ϕ. A key result in the theory of ODEs and generative modeling states that if two vector fields are close in L_p norm, their corresponding flows will also be close. Specifically, if $\|v_\phi - v^*\|_{L_2(p_t)}$ is small, then the resulting distributions p_1^ϕ and $p_1(x;G)$ will also be close in a suitable statistical distance (e.g., L_2 distance between densities, or integral probability metrics like MMD). This is formalized in Proposition 4.2.

For a sufficiently accurate approximation (i.e., ϵ is small in Proposition 4.2), the generated distribution p_1^ϕ will be arbitrarily close to the true target distribution $p_1(x;G)$.

Since $p_1(x; G)$ is by assumption causally consistent with G (as established in Theorem 3.1), the convergence of p_1^ϕ to $p_1(x; G)$ directly implies that the samples generated by integrating v_ϕ will also effectively preserve the causal structure dictated by G.

Thus, the optimization of the CCFM objective guarantees that the learned generative model accurately approximates the desired causally consistent data distribution.

□

A.4. Proof of Theorem 4.1 (Convergence of Interventional Flow Matching)

Theorem 4.1: Given an interventional DAG $G_{do(S)}$ and its corresponding post-interventional distribution $p_1^{do(S=s)}(x)$ (assumed to satisfy regularity conditions), the optimization of the I-CCFM Objective (4.1) ensures that the learned vector field $v_\phi(x, t)$ converges to the optimal interventional target vector field $v^*(x, t; G_{do(S)})$ in an L_2 sense. Consequently, the distribution p_1^ϕ generated by integrating v_ϕ from p_0 converges to $p_1^{do(S=s)}(x)$, thereby allowing for accurate generation of interventional data.

Proof:

The Interventional Causally Constrained Flow Matching (I-CCFM) objective is defined as:

$$\min_\phi \mathcal{L}_{\text{interventional}}(\phi) = \mathbb{E}_{t\sim U(0,1), x_0\sim p_0, x_1\sim p_1^{do(S=s)}(x)} \left[\|v_\phi((1-t)x_0 + tx_1, t) - (x_1 - x_0)\|_2^2 \right].$$

The crucial aspect of this objective is that the target samples x_1 are now explicitly drawn from the post-interventional distribution $p_1^{do(S=s)}(x)$. This distribution is itself derived from the original causal graph G via the *do*-calculus, specifically by the truncated product formula:

$$p_1^{do(S=s)}(x) = \left(\prod_{X_j \in V\setminus S} P(X_j | Pa_G(X_j)) \right) \cdot \mathbb{I}(x_S = s),$$

where $\mathbb{I}(x_S = s)$ is an indicator function ensuring that the intervened variables take on

their fixed values s. This definition implies that $p_1^{do(S=s)}(x)$ is causally consistent with the manipulated graph $G_{do(S)}$.

Let $x_t = (1-t)x_0 + tx_1$. Similar to the proof of Theorem 3.2, the optimal target vector field $v^*(x_t, t; G_{do(S)})$ for this specific path and marginal distributions p_0 and $p_1^{do(S=s)}(x)$ is given by the conditional expectation:

$$v^*(x_t, t; G_{do(S)}) = \mathbb{E}_{x_0 \sim p_0, x_1 \sim p_1^{do(S=s)}(x)} \left[(x_1 - x_0) \middle| x_t = (1-t)x_0 + tx_1 \right].$$

The existence and uniqueness of this optimal interventional target vector field follow directly from the arguments laid out in the proof of Theorem 3.1, merely replacing $p_1(x; G)$ with $p_1^{do(S=s)}(x)$. Since $p_1^{do(S=s)}(x)$ is assumed to satisfy regularity conditions (e.g., smooth density, finite moments), $v^*(x_t, t; G_{do(S)})$ will also be well-defined and unique for the chosen linear path.

The I-CCFM objective can then be re-written as an L_2 regression problem:

$$\mathcal{L}_{\text{interventional}}(\phi) = \mathbb{E}_{t \sim U(0,1), x_t \sim p_t^{do(S=s)}} \left[\| v_\phi(x_t, t) - v^*(x_t, t; G_{do(S)}) \|_2^2 \right],$$

where $p_t^{do(S=s)}$ is the marginal distribution of x_t along the path defined by $x_0 \sim p_0$ and $x_1 \sim p_1^{do(S=s)}(x)$.

1. Convergence of v_ϕ to $v^*(x, t; G_{do(S)})$: To prove the convergence of v_ϕ to $v^*(x, t; G_{do(S)})$, we rely on standard assumptions commonly made in the analysis of neural network training and Flow Matching models:

Universal Approximation Capability: The neural network model v_ϕ is assumed to have sufficient capacity (e.g., adequate number of layers and neurons) to act as a universal approximator. This implies that v_ϕ can approximate the true, possibly complex, target vector field $v^*(x, t; G_{do(S)})$ to an arbitrary degree of accuracy.

Regularity of Target Vector Field: The optimal interventional vector field $v^*(x, t; G_{do(S)})$ is assumed to be a sufficiently regular function (e.g., continuous, Lipschitz continuous) over the domain of interest. This holds if p_0 and $p_1^{do(S=s)}(x)$ are well-behaved distributions.

Sufficient Training Data: The training process has access to a sufficiently large

number of independent and identically distributed (i.i.d.) samples (x_0, x_1) where $x_0 \sim p_0$ and $x_1 \sim p_1^{do(S=s)}(x)$. This ensures that the empirical expectation used in stochastic optimization accurately approximates the true expectation in the objective function.

Convergent Optimization Algorithm: The optimization algorithm (e.g., stochastic gradient descent or its variants) is assumed to converge to the minimum of the objective function, given appropriate learning rate schedules and batch sizes.

Under these standard conditions, the minimization of the L_2 objective function $\mathcal{L}_{\text{interventional}}(\phi)$ will force the learned vector field $v_\phi(x,t)$ to converge to the true optimal interventional target vector field $v^*(x,t;G_{do(S)})$ in an L_2 sense with respect to the path distribution $p_t^{do(S=s)}$. That is:

$$\|v_\phi - v^*\|^2_{L_2(p_t^{do(S=s)})} = \mathbb{E}_{t\sim U(0,1), x_t\sim p_t^{do(S=s)}}\left[\|v_\phi(x_t,t) - v^*(x_t,t;G_{do(S)})\|_2^2\right] \to 0,$$

as the optimization converges to its minimum.

2. Consequent Convergence of p_1^ϕ to $p_1^{do(S=s)}(x)$: The convergence of the learned vector field v_ϕ to the true target vector field $v^*(x,t;G_{do(S)})$ directly implies the convergence of the probability distribution generated by integrating v_ϕ to the target distribution $p_1^{do(S=s)}(x)$. If we denote the flow generated by v_ϕ starting from $X_0^\phi \sim p_0$ as $\frac{dX_t^\phi}{dt} = v_\phi(X_t^\phi, t)$, then its distribution at $t=1$ is $p_1^\phi(x)$. Similarly, let $p_t^*(x)$ be the distribution generated by the optimal vector field $v^*(x,t;G_{do(S)})$, with $p_1^*(x) = p_1^{do(S=s)}(x)$.

The relationship between the proximity of vector fields and the proximity of their generated distributions is well-established in the theory of ODEs. If $\|v_\phi - v^*\|_{L_2(p_t^{do(S=s)})}$ is small, then the resulting distributions $p_1^\phi(x)$ and $p_1^{do(S=s)}(x)$ will also be close in a relevant statistical distance (e.g., total variation distance, L_2 distance between densities, or Maximum Mean Discrepancy (MMD)). Proposition 4.2 further quantifies this relationship.

Since the target distribution $p_1^{do(S=s)}(x)$ is, by its very definition, causally consistent with the interventional graph $G_{do(S)}$, the convergence of $p_1^\phi(x)$ to $p_1^{do(S=s)}(x)$ ensures that the samples generated by the I-CCFM framework accurately reflect the post-interventional causal relationships.

Thus, the optimization of the I-CCFM objective successfully trains a generative model that can accurately simulate the effects of specific causal interventions, providing a powerful tool for policy analysis.

□

8.1 A.5. Proof of Proposition 4.2 (Approximation Error Bound)

Proposition 4.2: Let v_ϕ be the learned vector field and v^* be the optimal target vector field for a given intervention $do(S = s)$. Assume v^* is Lipschitz continuous with constant L_{v^*}. If $\|v_\phi - v^*\|_{L_2(p_t)} \leq \epsilon$ for some small $\epsilon > 0$, then there exists a constant C (dependent on the time horizon $T = 1$ and L_{v^*}) such that the L_2 distance between the generated density p_1^ϕ and the true interventional density $p_1^{do(S=s)}$ is bounded by:

$$\|p_1^\phi - p_1^{do(S=s)}\|_{L_2} \leq C \cdot \epsilon.$$

Furthermore, the value of ϵ is influenced by:

1. Network Capacity: The universal approximation capabilities of neural networks imply that ϵ decreases as the capacity (e.g., number of layers, neurons) of v_ϕ increases, assuming sufficient data.

2. Causal Graph Complexity: The complexity of $p_1^{do(S=s)}$ itself, which can be influenced by the graph structure (e.g., maximum in-degree, treewidth), affects the difficulty of approximation. Interventions on variables with many children or in densely connected parts of the graph might lead to more complex conditional distributions, potentially increasing ϵ for a fixed network capacity.

Proof:

Let X_t^* denote the true trajectory generated by the optimal vector field $v^*(x_t, t; G_{do(S)})$ starting from $X_0 \sim p_0$:

$$\frac{dX_t^*}{dt} = v^*(X_t^*, t; G_{do(S)}), \quad X_0^* = X_0.$$

The distribution of X_1^* at $t = 1$ is $p_1^{do(S=s)}(x)$.

Let X_t^ϕ denote the trajectory generated by the learned vector field $v_\phi(x_t, t)$ starting from the same initial condition $X_0 \sim p_0$:

$$\frac{dX_t^\phi}{dt} = v_\phi(X_t^\phi, t), \quad X_0^\phi = X_0.$$

The distribution of X_1^ϕ at $t = 1$ is $p_1^\phi(x)$.

1. Bounding the Divergence of Trajectories (using Gronwall's Inequality): Consider the difference between the true and learned trajectories starting from the same X_0. For any $t \in [0, 1]$, we have:

$$X_t^\phi - X_t^* = \int_0^t [v_\phi(X_s^\phi, s) - v^*(X_s^*, s)]ds.$$

We can rewrite the integrand:

$$v_\phi(X_s^\phi, s) - v^*(X_s^*, s) = [v_\phi(X_s^\phi, s) - v^*(X_s^\phi, s)] + [v^*(X_s^\phi, s) - v^*(X_s^*, s)].$$

Taking the L_2 norm and applying the triangle inequality:

$$\|X_t^\phi - X_t^*\|_2 \leq \int_0^t \|v_\phi(X_s^\phi, s) - v^*(X_s^\phi, s)\|_2 ds + \int_0^t \|v^*(X_s^\phi, s) - v^*(X_s^*, s)\|_2 ds.$$

By the Lipschitz continuity of v^* with constant L_{v^*} (i.e., $\|v^*(x, t) - v^*(y, t)\|_2 \leq L_{v^*}\|x - y\|_2$):

$$\|X_t^\phi - X_t^*\|_2 \leq \int_0^t \|v_\phi(X_s^\phi, s) - v^*(X_s^\phi, s)\|_2 ds + L_{v^*} \int_0^t \|X_s^\phi - X_s^*\|_2 ds.$$

Let $e(t) = \|X_t^\phi - X_t^*\|_2$. Then, for any path from X_0:

$$e(t) \leq \int_0^t \|v_\phi(X_s^\phi, s) - v^*(X_s^\phi, s)\|_2 ds + L_{v^*} \int_0^t e(s)ds.$$

By Gronwall's Inequality, for the deterministic version:

$$e(t) \leq \left(\int_0^t \|v_\phi(X_s^\phi, s) - v^*(X_s^\phi, s)\|_2 ds \right) e^{L_{v^*} t}.$$

Averaging over the initial samples $X_0 \sim p_0$ and the time $t \sim U(0,1)$, we have that the expected L_2 difference between trajectories is bounded. More precisely, for the expected value of the L_2 difference at $t = 1$:

$$\mathbb{E}[\|X_1^\phi - X_1^*\|_2] \leq C_1 \mathbb{E}_{t \sim U(0,1), X_t^\phi \sim p_t^\phi}[\|v_\phi(X_t^\phi, t) - v^*(X_t^\phi, t)\|_2],$$

where $C_1 = e^{L_{v^*} T}$ (with $T = 1$ in our case). Note that the expectation in the theorem's assumption is over p_t, the true path distribution from v^*. If p_t^ϕ is sufficiently close to p_t, then this bound holds. More formally, we can bound the distance using the given condition. We are given $\|v_\phi - v^*\|_{L_2(p_t)} \leq \epsilon$, which means $\mathbb{E}_{t \sim U(0,1), x_t \sim p_t}[\|v_\phi(x_t, t) - v^*(x_t, t; G_{do(S)})\|_2^2] = \epsilon^2$.

2. Relate Trajectory Error to Distribution Error: The closeness of trajectories translates to closeness of distributions. For a smooth transformation governed by an ODE, if the vector fields are close, their pushforward measures will also be close. A known result (e.g., from density estimation or optimal transport literature for flow-based models) relates the L_2 distance between densities to the L_2 distance between their corresponding vector fields. For continuous flows that pushforward p_0 to p_1, if $\|v_\phi - v^*\|_{L_2(p_t)} \leq \epsilon$, then there exists a constant C_2 such that the L_2 distance between the final densities p_1^ϕ and $p_1^{do(S=s)}$ is bounded:

$$\|p_1^\phi - p_1^{do(S=s)}\|_{L_2} \leq C_2 \cdot \|v_\phi - v^*\|_{L_2(p_t)},$$

where C_2 depends on the time horizon T, the Lipschitz constant L_{v^*}, and potentially other properties like the inverse Lipschitz constant of the flow or the regularity of the densities. Substituting the given condition, we obtain:

$$\|p_1^\phi - p_1^{do(S=s)}\|_{L_2} \leq C_2 \cdot \epsilon.$$

Combining C_1 and C_2 into a single constant $C = C_1 C_2$ gives the desired bound:

$$\|p_1^{\phi} - p_1^{do(S=s)}\|_{L_2} \leq C \cdot \epsilon.$$

This establishes the first part of the proposition.

3. Factors Influencing ϵ: The error $\epsilon = \|v_\phi - v^*\|_{L_2(p_t)}$ is the L_2 approximation error of the neural network v_ϕ to the true target vector field v^*. This error is influenced by two primary factors:

Network Capacity: Neural networks are known as universal approximators. This means that, given sufficient depth and width (i.e., capacity), a neural network can approximate any continuous function to an arbitrary degree of accuracy over a compact set. Therefore, as the capacity of v_ϕ increases, its ability to approximate the true $v^*(x_t, t; G_{do(S)})$ improves, leading to a smaller ϵ. This implicitly assumes that the training process converges to a global optimum and that there is sufficient data to constrain the approximation.

Causal Graph Complexity: The complexity of the true target vector field $v^*(x_t, t; G_{do(S)})$ is inherently tied to the complexity of the post-interventional distribution $p_1^{do(S=s)}(x)$, which in turn is determined by the structure of the causal graph G and the intervention $do(S = s)$.

If G is simple (e.g., a chain or a star graph) or the intervention is simple (e.g., affecting only an exogenous variable), the conditional distributions $P(X_j|Pa_G(X_j))$ and thus $p_1^{do(S=s)}(x)$ might be relatively simple functions. In such cases, v^* would also be relatively simple, allowing v_ϕ to approximate it with a smaller ϵ using less capacity.

Conversely, if G is complex (e.g., dense, high maximum in-degree, or involves many non-linear relationships), the conditional distributions can be highly intricate. For example, variables with numerous parents will have complex conditional dependencies. When an intervention occurs, the re-factored distribution $p_1^{do(S=s)}(x)$ can become significantly more complex, requiring v^* to capture these intricate relationships. A more complex v^* necessitates a higher-capacity neural network to achieve a given ϵ, or for a fixed network capacity, the approximation error ϵ will be larger. Properties like the treewidth of the

graph, or the number and nature of causal paths, directly impact the complexity of the underlying causal model and thus the functional form of v^*.

This proposition underscores the theoretical relationship between the learnability of the vector field and the accuracy of the generated interventional distributions, providing guidance on the required model complexity for accurate GCMs.

□