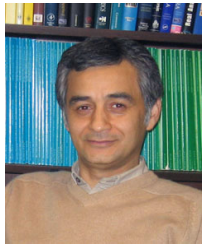
NODAGS-Flow: Nonlinear Cyclic Causal Structure Learning


Murali G Sethuraman[1]


Romain Lopez[2,3]


Rahul Mohan[2]


Faramarz Fekri[1]


Tommaso Biancalani[2]


Jan-Christian Hütter[2]

[1]School of ECE, Georgia Insititute of Technology

[2]Division of Research and Early Development, Genentech

[3]Department of Genetics, Stanford University

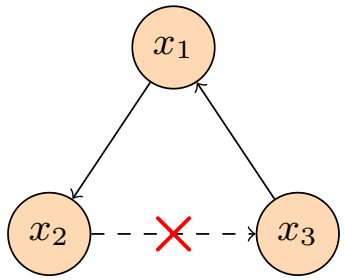
Georgia Institute of Technology

Genentech
A Member of the Roche Group

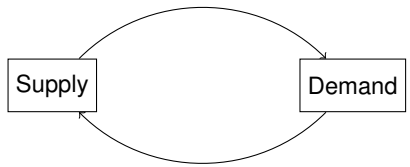
Motivation

- Causal understanding of real-world systems is crucial for prediction under unseen interventions.

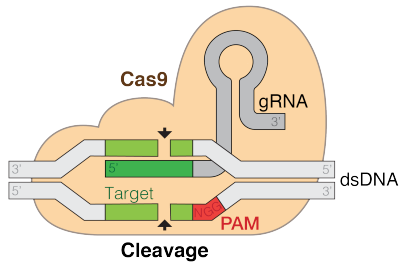
- With a few notable exceptions, majority of **causal discovery** (CD) methods rely on structure being a **directed acyclic graph** (DAG).



- DAG Assumption:
  - ✓ Regularize the search space.
  - × Not realistic in practice.



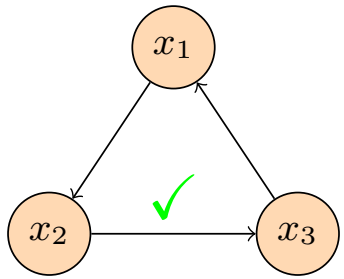
- Recent advances in biological assays based on CRISPR-Cas9 allow for large scale **interventions** on gene networks, enabling testing on large real-world data sets.



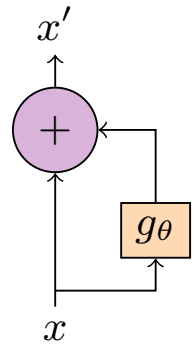
The image was taken from: `https://en.wikipedia.org/wiki/CRISPR_gene_editing`.

Contributions

- Novel framework for causal discovery that allows for **cycles** and flexible **nonlinear causal relations**.



- **Maximum likelihood estimation** (MLE) based graph recovery utilizing **contractive residual flows**.



- We provide a **differentiable**, **consistent** causal graph estimator that can handle both **observational** and **interventional** data.
- The proposed optimization program **avoids any parameter tuning** for constrained optimization as in most algorithms of this class.
- Through experiments on **synthetic** and **real-world** data, we showcase the benefits of the proposed method.

Problem Setup

Let $G = (V, E)$ be a **directed cyclic** graph that represents a discrete dynamical system with observations at time t, $x^{(t)} = (x_1^{(t)}, \ldots, x_d^{(t)})$.

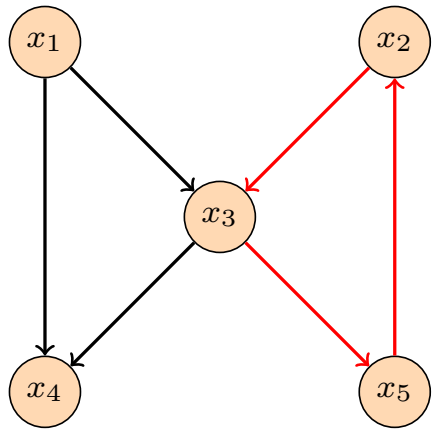
Structural Equation Model (SEM)

At time t, The causal relations can be stated as follows

$$x_i^{(t)} = f_i\left(x_{\text{pa}(i)}^{(t-1)}\right) + \varepsilon_i, \quad i = 1, \ldots, d.$$

where $\varepsilon_i \sim p_E(\varepsilon)$. Upon **vectorization** we get:

$$x^{(t)} = f\left(x^{(t-1)}\right) + \varepsilon, \quad f = (f_1, \ldots, f_d).$$



At equilibrium: (we consider all observed data to be from equilibrium state)

$$x^* = f(x^*) + \varepsilon \implies \varepsilon = (\text{id} - f)(x^*) \quad \text{where} \quad \text{id}(x) = x.$$

Assumptions

- No *confounders*
- p_E is known
- $(\text{id} - f)$ differentiable
- $(\text{id} - f)^{-1}$ exists, *differentiable*

Differentiable Causal Discovery

Given a directed graph (DG) G, a set of observations at steady state $\{x^i\}_{i=1}^n$.

DCDI[1] **(observational)**: $\boxed{G \in \text{DAGs}}$

The graph has to be a DAG for likelihood to be computed correctly

$$\max_\theta \underbrace{\sum_{j=1}^{n} \boxed{\sum_{i=1}^{d} \log p_\theta \left(x_i^j | x_{\text{pa}(i)}^j \right)}}_{\text{data likelihood}} - \overbrace{\lambda \Omega(\theta)}^{\text{sparsity regularizer}}$$

Expensive to compute!

$$\text{s.t} \quad \boxed{\underbrace{\text{Tr} e^{A_\theta} - d = 0}_{\text{DAG constraint}}}$$

Requires augmented lagrangian based optimization

NODAGS-Flow: $\boxed{G \in \text{DGs}}$

Potentially expensive!

$$\max_\theta \underbrace{\sum_{j=1}^{n} \boxed{\log p_X(x^j)}}_{\text{data likelihood}} - \overbrace{\lambda \Omega(\theta)}^{\text{sparsity regularizer}}$$

[1] Philippe Brouillard et al. "Differentiable causal discovery from interventional data". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 21865–21877.

NODAGS-Flow: Data Likelihood

Density of x. Given the **structural equations** and the **assumptions** the density of the observations x is given by

$$p_x(x) = p_E\Big((\mathrm{id} - f)(x)\Big)\Big|\det J_{(\mathrm{id}-f)}(x)\Big|$$

where $J_{(\mathrm{id}-f)}(x)$ is the Jacobian operator of $(\mathrm{id} - f)$ at x.

Data likelihood. Given a set of observations $\{x^j\}_{j=1}^n$, the log-**data likelihood** is given by

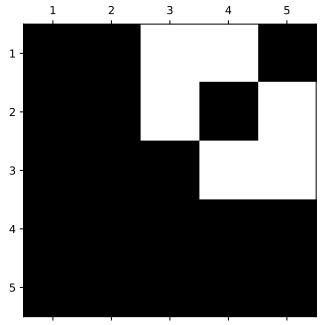
Expensive when computed naïvely!

$$\mathcal{L}(f, \{x^j\}_{j=1}^n) = \sum_{j=1}^{n} \log P_E\Big((\mathrm{id} - f)(x^j)\Big) + \underbrace{\boxed{\log\Big|\det J_{(\mathrm{id}-f)}(x^j)\Big|}}_{\text{Log-det-Jacobian}}$$

NODAGS-Flow: Causal Mechanism

The causal mechanism f is modeled using **contractive neural networks** (NN), with the parent-child relations encoded by an **input dependency mask**[2] $M \in \{0,1\}^{d \times d}$.

Also removes self loops!



Dependency mask

Definition (Contractive function)

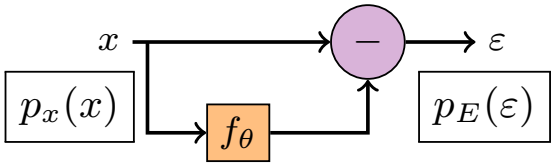
A function $f : \mathbb{R}^d \to \mathbb{R}^d$ is contractive if $\exists\, L < 1$ such that for $z_1, z_2 \in \mathbb{R}^d$,

$$\|f(z_1) - f(z_2)\| \leq L\|z_1 - z_2\|.$$

$$f_i(x) = \mathsf{NN}(x \odot M_{i,*})$$

Contractivity is enforced by rescaling NN weights.

The map $x \mapsto \varepsilon = (\mathsf{id} - f_\theta)$ forms a **residual normalizing flow** network.



$$p_X(x) = p_E\Big((\mathsf{id} - f_\theta)(x)\Big)\,\big|\det J_{(\mathsf{id}-f_\theta)}(x)\big|$$

Contractive $f \implies (\mathsf{id} - f)^{-1}$ exists. Inverse computed using **fixed point iteration**.

[2] Eric Jang, Shixiang Gu, and Ben Poole. "Categorical Reparameterization with Gumbel-Softmax". In: *International Conference on Learning Representations*. 2017.

NODAGS-Flow: Computing Log-det-Jacobian

- **Power series expansion**

$$\log|\det J_{(\mathrm{id}-f)}(x)| = \log|\det\big(\boldsymbol{I} - J_f(x)\big)| = -\sum_{k=1}^{\infty}\frac{1}{k}\mathsf{Tr}\Big\{J_f^k(x)\Big\}$$

The power series converges when f is contractive.

- **Hutchinson Trace estimator**[3]: $\mathsf{Tr}\big\{J_f^k(x)\big\} = \mathbb{E}_w[w^\top J_f^k(x)w]$, where $\mathbb{E}w = 0$, and $\mathbb{E}w^2 = 1$.

- **Removing bias**[4]: sample $n \sim P(N)$ and re-weight.

$$\log|\det J_{(\mathrm{id}-f)}(x)| = -\mathop{\mathbb{E}}_{n,w}\left[\sum_{k=1}^{n}\frac{w^\top J_f^k(x)w}{k\cdot P(N\geq k)}\right].$$

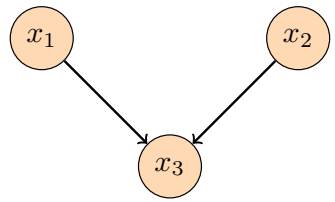
[3] Jens Behrmann et al. "Invertible residual networks". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 573–582.

[4] Ricky TQ Chen et al. "Residual flows for invertible generative modeling". In: *Advances in Neural Information Processing Systems* 32 (2019).

Extending to Non-contractive DAGs

Although contractivity is *sufficient* for invertibility of $(\mathrm{id} - f)$, it is **not necessary**.

For the case of DAGs, f need not be contractive. To tackle this issue, we add preconditioning terms Λ to the model.



$$\boldsymbol{W} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}$$

$$L_W = \|\boldsymbol{W}\| = 2.23$$

Non-contractive DAG.

Proposition

Let (G, f) represent a causal DAG and its causal mechanism. If f is a non-contractive function, then there exists $\tilde{f}$ of the form $\tilde{f} = \Lambda \circ f \circ \Lambda^{-1}$, where Λ denotes multiplication with a diagonal matrix with positive diagonal entries such that $\tilde{f}$ is contractive.

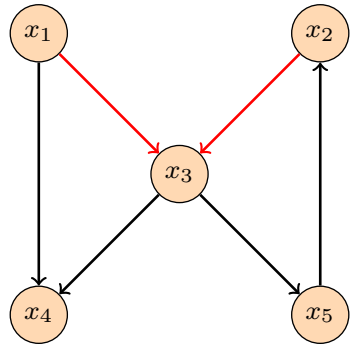
Final objective: (**Amenable to stochastic gradient based solvers**)

$$\max_{\theta, \Lambda} \mathop{\mathbb{E}}_{\boldsymbol{M}' \sim \boldsymbol{M}_\phi} \mathcal{L}(\theta, \Lambda^{-1} \circ f_\theta \circ \Lambda, \boldsymbol{M}') - \lambda \mathop{\mathbb{E}}_{\boldsymbol{M}' \sim \boldsymbol{M}_\phi} [\|\boldsymbol{M}'\|_1]. \qquad (1)$$

Modeling Interventions

We only consider **surgical** (perfect) interventions. For an experimental setting $\mathcal{E} = (\mathcal{I}, \mathcal{U})$, we have the following SEM,

$$x_i = \tilde{\varepsilon}_i, \quad i \in \mathcal{I};$$
$$x_i = f(x_{\mathsf{pa}(i)}) + \varepsilon_i, \quad i \in \mathcal{U}.$$



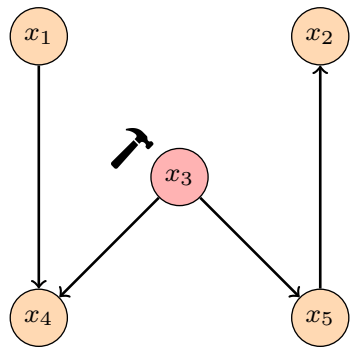
$\mathcal{I} = \emptyset$

Upon vectorization we then get,

$$x = \boldsymbol{U}\big(f(x) + \varepsilon\big) + \tilde{\varepsilon},$$

where

- $\tilde{\varepsilon}$ - value of intervened nodes.
- $\boldsymbol{U}$ - diagonal matrix with 1 corresponding to $\mathcal{U}$.



$\mathcal{I} = \{x_3\}$

Modified density. The density of x under the intervention $\mathcal{I}$ is given by

$$p_x(x) \propto p_E\Big(\big[(\mathsf{id} - \boldsymbol{U}f)(x)\big]_{\mathcal{U}}\Big)\Big|\det J_{(\mathsf{id}-\boldsymbol{U}f)}(x)\Big|.$$

Consistency under Linear SEM

Restricting to the case where SEM is **linear** and disturbance is **Gaussian**. That is,

Linear Gaussian SEM: $x = B^\top x + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Omega).$

Then the graph estimated by MLE is identifiable to the ground truth graph up to a notion of **quasi-equivalence**[5] extended to allow interventions.

Theorem

*Under some mild assumptions, the global minimizer of (1) with a suitably chosen λ outputs $\hat{G} \cong_{\mathcal{I}} G$ (**interventionally quasi-equivalent**) asymptotically, where G is the ground truth graph.*

[5] AmirEmad Ghassami et al. "Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 3494–3504.

Experiments - Gaussian Structural Causal Models

Baselines

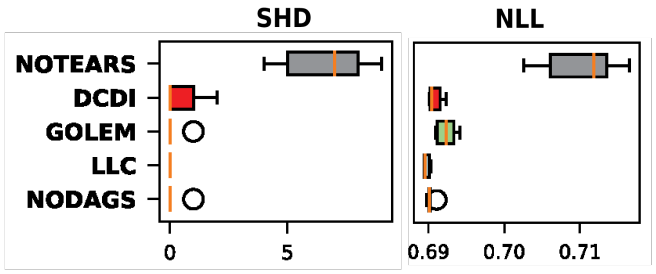
Baseline	SEM	Graph
NOTEARS	Linear	DAG
DCDI	Nonlinear	DAG
GOLEM	Linear	DAG
LLC	Linear	Cyclic

Experiments

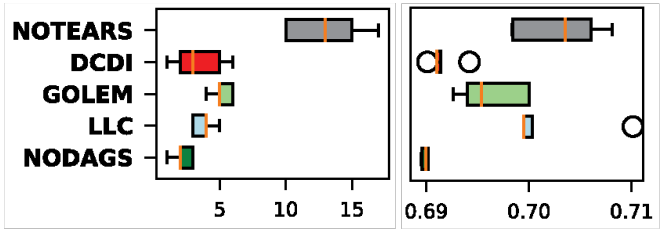
Setting	SEM	Graph
int-dag-lin	Linear	DAG
int-dag-nonlin	Nonlinear	DAG
int-cyc-lin	Linear	Cyclic
int-cyc-nonlin	Nonlinear	Cyclic

NODAGS-Flow beats all the baselines on nonlinear cyclic SEMs and is competitive with the baseline in other settings.

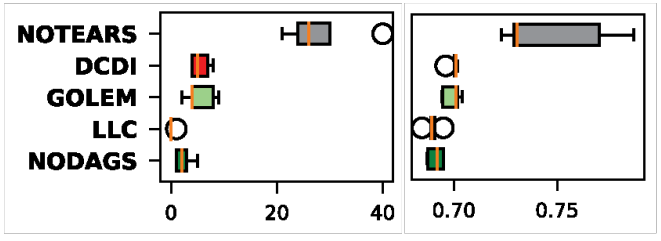
Interventional-DAG-Linear SEM (*int-dag-lin*)



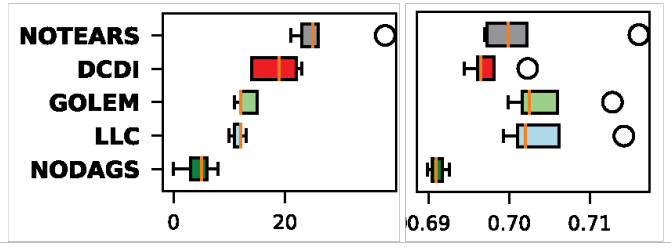
Interventional-DAG-Nonlinear SEM (*int-dag-nonlin*)



Interventional-Cyclic-Linear SEM (*int-cyc-lin*)



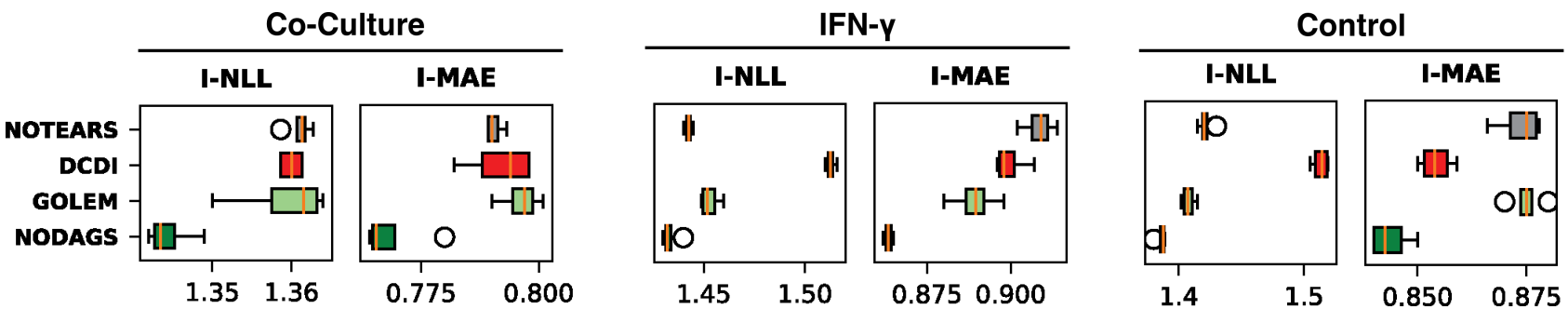
Interventional-Cyclic-Nonlinear SEM (*int-cyc-nonlin*)



Experiments - Gene Perturbation dataset

Perturb-seq dataset

- The data was taken from Frangieh et al (2021)[6].
- Contains gene expressions taken from 218,331 melanoma cells split over (1) control (57,627 cells) , (2) co-culture (73,114 cells), and (3) interferon (IFN)-γ (87,590 cells).
- Due to practical limitations we restrict to a subset of 61 genes.



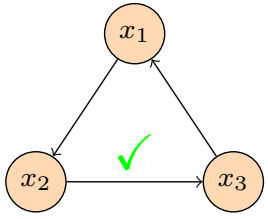
NODAGS-Flow beats all the baselines over all the datasets.

[6] Chris J Frangieh et al. "Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion". In: *Nature genetics* 53.3 (2021), pp. 332–341.

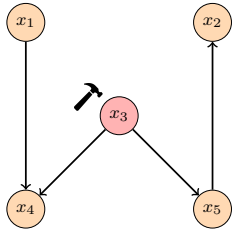
Conclusion

- **NODAGS-Flow**: Differentiable nonlinear causal graph recovery.

Allows cycles



Handles interventions



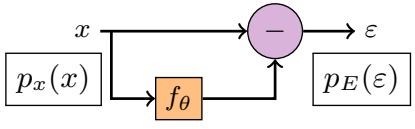
Maximum likelihood estimator

$$\max_{\theta,\Lambda} \mathop{\mathbb{E}}_{M' \sim M_\phi} \mathcal{L}(\theta, \Lambda^{-1} \circ f_\theta \circ \Lambda, M') - \lambda \mathop{\mathbb{E}}_{M' \sim M_\phi} [\|M'\|_1].$$

No parameter tuning for constrained optimization

- Using **contractive NNs** and **residual flows**, the likelihood is efficiently computed.

Residual Flow

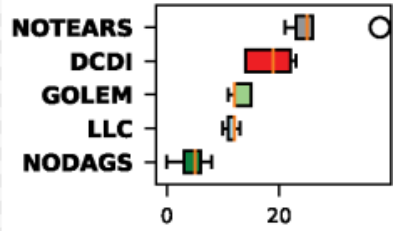


Efficient log-det-Jacobian estimator

$$\log|\det J_{(\mathrm{id}-f)}(x)| = -\mathop{\mathbb{E}}_{n,w}\left[\sum_{k=1}^{n} \frac{w^\top J_f^k(x) w}{k \cdot P(N \geq k)}\right]$$

- Showcased **improved recovery performance** on synthetic and real-world data sets.

SHD - cyclic nonline (synthetic)



NLL - Real-world data