Information-Geometric Neural Granger Causality

Pauline Bourigault P.BOURIGAULT22@IMPERIAL.AC.UK
Imperial College London, United Kingdom

Danilo Mandic D.MANDIC@IMPERIAL.AC.UK
Imperial College London, United Kingdom

Abstract

Discovering causal relationships from time series data is fundamental across scientific domains. While neural networks have advanced time series analysis, neural approaches to Granger causality lack theoretical foundations—particularly recent methods that achieve strong empirical performance without explanatory theory. We introduce Information-Geometric Neural Granger Causality (IG-NGC), a framework revealing how neural networks discover causal relationships through learning statistical manifolds. Our key insight is that causality manifests as directional information flow measured by the Fisher metric on these manifolds. This geometric perspective provides a theoretical framework that connects neural causality methods: we prove that causal influences emerge as information-geometric properties and demonstrate how existing approaches can be interpreted as approximations of our Information-Geometric Granger Causality (IGGC) measure, with exact equivalence established under Gaussian assumptions. Our framework yields theoretical guarantees including consistency, finite-sample complexity bounds, and scale consistency. Experiments on synthetic and real-world datasets confirm our theoretical predictions, transforming neural causality from empirical techniques into theoretically grounded methodologies.

1. Introduction

Discovering causal relationships from time series data is fundamental across scientific domains, from neuroscience [20] to economics [10] and climate research [18]. While neural networks have significantly advanced time series analysis, neural approaches to Granger causality lack theoretical foundations—particularly recent methods that achieve strong empirical performance without explanatory theory. Neural Granger causality methods face two critical limitations: (1) computational inefficiency from training separate models for each target variable, and (2) architectural constraints from first-layer weight sparsity assumptions. The recent Jacobian Regularizer-based Neural Granger Causality (JRNGC) [26] provides a practical solution by using a single model with Jacobian regularization, but offers no theoretical explanation for why this approach works.

We bridge this gap by introducing Information-Geometric Neural Granger Causality (IG-NGC), a framework revealing how neural networks discover causal relationships through learning statistical manifolds. Our key insight is that causality manifests as directional information flow measured by the Fisher metric on these manifolds. This geometric perspective provides a unifying theory for neural causality methods: we prove that causal influences emerge as information-geometric properties and demonstrate how existing approaches, including JRNGC [26], implicitly approximate our fundamental Information-Geometric Granger Causality (IGGC) measure. Our contributions include: (1) establishing the first information-geometric theory of neural Granger causality, (2) proving that existing methods approximate more fundamental information-geometric quantities,

© P. Bourigault & D. Mandic.

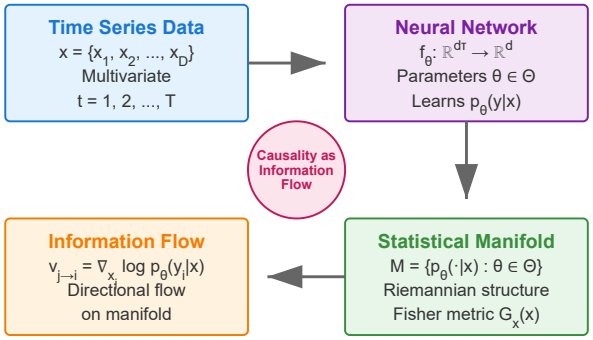

Figure 1: **Information-Geometric Neural Granger Causality (IG-NGC) framework.** Time series data $\mathbf{x} = \{x^1, x^2, \ldots, x^D\}$ with D dimensions and time indices $t = 1, 2, \ldots, T$ is processed through a neural network $f_\theta : \mathbb{R}^{d\tau} \to \mathbb{R}^d$ with parameters $\theta \in \Theta$ that learns conditional distributions $p_\theta(y|x)$. This induces a statistical manifold $\mathcal{M} = \{p_\theta(\cdot|x) : \theta \in \Theta\}$ equipped with the Riemannian structure from the Fisher metric $G_x(x)$. Causal relationships manifest as directional information flow $v_{j\to i} = \nabla_{x_j} \log p_\theta(y_i|x)$ on this manifold. Our framework unifies existing neural causality methods—component-wise models (cMLP, cLSTM), Jacobian Regularizer-based methods (JRNGC), additive structure methods (NAVAR), and attention-based approaches (TCDF)—as special cases that implicitly approximate the fundamental information-geometric quantities.

(3) deriving theoretical guarantees including consistency and finite-sample complexity bounds, and (4) explaining why different architectural choices can successfully capture causality through their induced manifold geometry. Through experiments on synthetic and real-world datasets, we demonstrate that our theoretical predictions match empirical behavior, transforming neural causality from empirical techniques into theoretically grounded methodologies.

2. Background and Related Work

Granger Causality Granger causality [9] formalizes the intuitive notion that causes precede effects by testing whether past values of one time series improve predictions of another. For multivariate time series $\mathbf{x} = \{x^1, x^2, \ldots, x^D\}$ with D dimensions, we consider a maximum lag τ such that predictions use past values from $t - \tau$ to $t - 1$. Variable j Granger-causes variable i if $P(x_i(t) \mid \mathbf{x}(< t)) \neq P(x_i(t) \mid \mathbf{x}_{-j}(< t))$, where $\mathbf{x}_{-j}$ denotes all variables except j, and $\mathbf{x}(< t) = \{x(t-\tau), \ldots, x(t-1)\}$ represents past values within the lag window.

Neural Granger Causality Recent work leverages neural networks for Granger causality, trading traditional statistical guarantees for empirical effectiveness. Tank *et al.* [22] introduced component-wise models (cMLP, cLSTM) using sparsity constraints on first-layer weights, given by

$$\hat{x}_i(t) = f_i\Big(\mathbf{x}(t-\tau : t-1); \theta_i\Big), \quad \mathcal{L}_i = \text{MSE}_i + \lambda \left\| W_i^{(1)} \right\|_1 . \tag{1}$$

JRNGC [26] addresses computational inefficiency by using a single model with Jacobian regularization as

$$\mathcal{L} = \frac{1}{N} \sum_t \left\| \mathbf{x}_t - f\left(\mathbf{x}_{t-\tau:t-1}\right) \right\|^2 + \lambda \left\| J \right\|_F^2 , \tag{2}$$

where $J_{ij} = \partial f_i / \partial x_j$ is the input-output Jacobian.

Information Geometry Information geometry [2] studies statistical models as Riemannian manifolds equipped with the Fisher information metric. This geometric perspective has yielded insights in optimization (*e.g.*, natural gradient descent [1]), representation learning (*e.g.*, information bottleneck [23]), and probabilistic modeling (*e.g.*, normalizing flows [17]), but its application to causal discovery remains unexplored. See Appendix A for further discussion on related work.

3. Information-Geometric Framework

We now present our framework that helps revealing why neural networks can discover causal relationships, through the lens of information geometry. Our key insight is that neural networks implicitly learn statistical manifolds where causality manifests as information flow. The detailed proofs for this section can be found in Appendix B.

Definition 1 (Neural Statistical Manifold) *Let $f_\theta : \mathbb{R}^{d\tau} \to \mathbb{R}^d$ be a neural network with parameters $\theta \in \Theta \subseteq \mathbb{R}^k$. The statistical manifold is defined as $\mathcal{M} = \{p_\theta(\cdot \mid x) : \theta \in \Theta\}$, where p_θ represents the conditional distribution parametrized by f_θ.*

This manifold has a natural Riemannian structure induced by the Fisher Information.

Definition 2 (Fisher Information Metric) *The Riemannian metric on $\mathcal{M}$ induced by the Fisher Information Matrix in parameter space is given by*

$$G_{ij}(\theta) = \mathbb{E}_{p(x,y)} \left[\frac{\partial \log p_\theta(y \mid x)}{\partial \theta_i} \frac{\partial \log p_\theta(y \mid x)}{\partial \theta_j} \right]. \tag{3}$$

For causal analysis, we define the Fisher Information Metric in input space for the conditional distribution $p_\theta(y \mid x)$: $G_x(x) = \mathbb{E}_{y \mid x}\left[(\nabla_x \log p_\theta(y \mid x))(\nabla_x \log p_\theta(y \mid x))^T\right]$, where ∇_x denotes the gradient with respect to input x. The key insight is that causal relationships manifest as directional information flow on the manifold.

Proposition 1 (Causal Information Flow) *The causal influence from input x_j to output y_i is characterized by the gradient vector field as*

$$v_{j \to i}(x) = \nabla_{x_j} \log p_\theta(y_i \mid x). \tag{4}$$

The magnitude of this flow, measured by the Fisher metric, quantifies causal strength.

When variable j causes variable i, perturbations in j create measurable information flow toward i. This flow is naturally measured using the input-space Fisher metric, which accounts for the manifold's geometry and the network's transformation structure. Building on this geometric understanding, we formalize how causality is encoded in neural networks.

Definition 3 (Information-Geometric Granger Causality (IGGC)) *The IGGC from variable j to variable i at lag ℓ is defined as*

$$IGGC_{j \to i}^{(\ell)} = \mathbb{E}_{x \sim p(x)} \left[\left\| \nabla_{x_j^{t-\ell}} \log p_\theta\left(x_i^{t+1} \mid x^{t-\tau:t}\right) \right\|_{G_x(x)}^2 \right], \tag{5}$$

where $\| \cdot \|_{G_x(x)}$ is the norm induced by the input-space Fisher Information Matrix.

IGGC measures the expected squared magnitude of information flow from past values of j to future values of i, accounting for the geometry of the probability space.

4. Key Theoretical Results

Theorem 1 (Fundamental Properties of IGGC) *IGGC satisfies the following properties: (1) Non-negativity: $IGGC^{(\ell)}_{j\to i} \geq 0$; (2) Causal identification: $IGGC^{(\ell)}_{j\to i} = 0 \Leftrightarrow X_j^{t-\ell} \perp\!\!\!\perp X_i^{t+1} \mid \mathbf{X}_{-j}^{t-\tau:t}$; (3) Scale consistency: Under Gaussian conditional distributions $p_\theta(y|x) = \mathcal{N}(f_\theta(x), \sigma^2 I)$, if the noise variance scales as $\sigma^2 \to c\sigma^2$, then $IGGC^{(\ell)}_{j\to i} \to c^{-1} IGGC^{(\ell)}_{j\to i}$*

These properties establish IGGC as a theoretically sound measure of causality. Non-negativity follows from the positive semi-definiteness of the input-space Fisher metric. Under appropriate conditions, causal identification links IGGC to the conditional independence relationships that define Granger causality. Scale consistency ensures that causal relationships are detected consistently across different noise levels, with IGGC scaling as σ^{-2} under Gaussian assumptions. We can now establish the relation to JRNGC [26] via information geometry, explaining why Jacobian regularization works.

Theorem 2 (JRNGC as Approximation of IGGC) *Under Gaussian output distribution $p_\theta(y \mid x) = \mathcal{N}(f_\theta(x), \sigma^2 I)$ and assuming diagonal Fisher approximation (see Remark 1), the input-space Fisher matrix is $G_x(x) = \sigma^{-2} J(x)^T J(x)$, and we have*

$$IGGC^{(\ell)}_{j\to i} = \sigma^{-2}\mathbb{E}\left[\left(\frac{\partial f_{\theta,i}(x)}{\partial x_j^{t-\ell}}\right)^2\right] = \sigma^{-2} JRNGC^{(\ell)}_{j\to i}, \tag{6}$$

where $JRNGC^{(\ell)}_{j\to i} = \mathbb{E}\left[\left(\frac{\partial f_{\theta,i}(x)}{\partial x_j^{t-\ell}}\right)^2\right]$ is the Jacobian regularization term used in JRNGC.

Remark 1 (Diagonal Fisher Approximation) *Theorem 2 assumes the Fisher metric can be treated component-wise for scalar partial derivatives, corresponding to the diagonal Fisher approximation commonly used in practice. This assumption is reasonable when inputs have limited cross-correlations or when computational efficiency requires diagonal approximations.*

Theorem 2 reveals that JRNGC's Jacobian regularization equals the IGGC measure scaled by σ^{-2} under Gaussian assumptions. As the noise level $\sigma \to 0$ (deterministic limit), the relative difference between normalized versions of these measures vanishes, explaining JRNGC's empirical success.

Our information-geometric framework provides unified explanations for diverse neural approaches. For component-wise models (cMLP, cLSTM) [22], each model f_{θ_i} learns a separate manifold slice $\mathcal{M}_i = \{p_{\theta_i}(x_i^{t+1} \mid \mathbf{x}^{t-\tau:t}) : \theta_i \in \Theta_i\}$. The IGGC measure emerges within each slice, with the component-wise architecture preventing inter-variable information leakage that would occur with shared hidden layers. Specifically, for each target variable i, the gradient $\nabla_{x_j^{t-\ell}} \log p_{\theta_i}(x_i^{t+1} \mid \mathbf{x}^{t-\tau:t})$ captures the information flow from source j to target i without interference from other targets. Regarding additive methods (NAVAR) [6], the additive structure $f_{\theta,i}(x) = \sum_{j=1}^{d}\sum_{\ell=1}^{\tau} h_{ij}^{(\ell)}(x_j^{t-\ell})$ preserves interpretability through decomposition. Under our framework, the gradient decomposes as $\partial f_{\theta,i}(x)/\partial x_j^{t-\ell} = \partial h_{ij}^{(\ell)}(x_j^{t-\ell})/\partial x_j^{t-\ell}$, allowing IGGC to directly measure component-specific information flow. For attention-based methods (TCDF) [16], while not explicitly designed for information geometry, attention weights $\alpha_{ij}^{(\ell)}$ correlate with the magnitude of causal information flow $\|\nabla_{x_j^{t-\ell}} \log p_\theta(x_i^{t+1} \mid \mathbf{x}^{t-\tau:t})\|$, as both quantify the importance of input $x_j^{t-\ell}$ for predicting x_i^{t+1}.

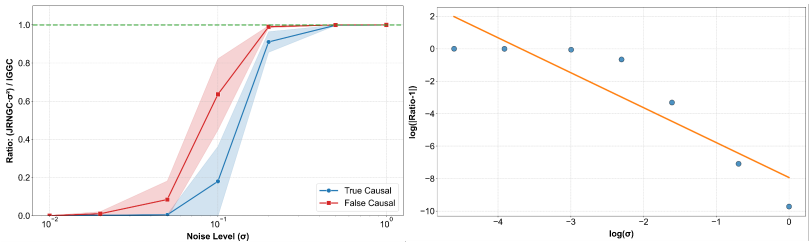

Figure 2: Theorem 2 validation, ratio convergence of JRNGC σ^2/IGGC to 1 as noise decreases. **Left**: Convergence by causal status (true causal vs. non-causal edges), where the ratio is computed as $\|\mathbf{J}\|_F^2\,\sigma^2/\text{tr}(\text{IGGC})$ with $\mathbf{J}$ the Jacobian matrix and IGGC the information-geometric causality matrix. **Right**: Log-log plot showing near-linear relationship between deviation from 1 and noise level.

Theorem 3 (Asymptotic Consistency) *Under identifiability conditions and appropriate regularization decay* $(\lambda_n \to 0, \lambda_n\sqrt{n} \to \infty)$*, the estimated causal graph converges as* $P(\hat{G}_n = G^*) \to 1$ *as* $n \to \infty$.

Theorem 4 (Finite-Sample Complexity) *Assume neural networks with bounded weights* $\|\theta\|_\infty \leq B$*, L-layer architecture with Lipschitz activation functions, and minimum detectable effect size* ε*. With probability* $\geq 1-\delta$*, IG-NGC recovers the true causal graph using*

$$n = O\left(\frac{d^2\tau\log(d/\delta)}{\varepsilon^2} + \frac{k\log(k/\delta)}{\varepsilon^2}\right) \tag{7}$$

samples, where k *is the number of network parameters.*

These results strengthen the statistical foundations of neural Granger causality.

5. Empirical Validation

Validating the Information-Geometric Framework We empirically validated Theorem 2, which posits that under Gaussian output distributions, JRNGC serves as an approximation of the IGGC measure. The theorem predicts that as noise level σ approaches zero, the ratio $\frac{\text{JRNGC}\sigma^2}{\text{IGGC}}$ should converge to 1. We tested this prediction using synthetic multivariate time series with known causal structure. For seven noise levels ($\sigma = 0.01$ to 1.0), Figure 2 confirms our theory: the ratio converges to 1 as noise decreases, with a near-linear relationship in log-log space between deviation and noise level ($R^2 = 0.825$). The negative slope indicates approximation error diminishes predictably as noise decreases.

Understanding Existing Methods Through IG-NGC We evaluated our IG-NGC framework against leading neural causality methods. Table 1 reveals two key findings: (1) IGGC closely approximates Jacobian-based summary causality (within ± 0.01), confirming Theorem 2's prediction that JRNGC's empirical success derives from approximating information-geometric quantities; and (2) performance degradation patterns follow theoretical expectations, with all methods excelling in low/medium-dimensional settings (VAR-50, Lorenz-96) while exhibiting anticipated precision challenges in high-dimensional spaces (VAR-100) and non-stationary systems (fMRI). The slight performance advantage of IGGC in highly nonlinear settings (Lorenz-96, $F = 40$) further supports

Table 1: Performance of JRNGC, IG-NGC and NAVAR

Dataset	AUROC				AUPRC			
	With Lag	With No Lag	IGGC	NAVAR(MLP)	With Lag	With No Lag	IGGC	NAVAR(MLP)
VAR(100, 5, 10)	**0.964**±0.007	0.879±0.012	0.878±0.012	0.887±0.023	0.673±0.031	0.745±0.024	0.744±0.024	**0.776**±0.040
VAR(50, 5, 10)	**0.999**±0.000	0.992±0.001	0.991±0.001	0.960±0.019	0.972±0.001	**0.977**±0.001	0.976±0.001	0.909±0.043
VAR(10, 3, 5)	**0.982**±0.019	0.960±0.040	0.949±0.052	0.976±0.131	0.889±0.112	**0.913**±0.091	0.867±0.126	0.936±0.086
Lorenz-96 (F=10)	**1.000**±0.000	**1.000**±0.000	**1.000**±0.000	0.993±0.004	0.996±0.003	**1.000**±0.000	**1.000**±0.000	0.986±0.008
Lorenz-96 (F=40)	**0.997**±0.002	0.974±0.011	0.976±0.012	0.900±0.021	0.963±0.016	**0.964**±0.016	**0.964**±0.019	0.828±0.052
fMRI	–	**0.816**±0.014	0.810±0.013	0.770±0.020	–	**0.618**±0.004	0.606±0.002	0.557±0.045

Table 2: Comparison of cMLP, cLSTM, and Component-wise IGGC

Dataset	AUROC			AUPRC		
	cMLP	cLSTM	Component IGGC	cMLP	cLSTM	Component IGGC
VAR(100, 5, 10)	**0.940**±0.013	0.845±0.045	0.870±0.007	**0.851**±0.051	0.606±0.102	0.729±0.015
VAR(10, 3, 5)	**0.978**±0.032	0.931±0.061	0.938±0.062	**0.923**±0.110	0.803±0.162	0.849±0.151
Lorenz-96 (F=10)	0.997±0.006	0.974±0.028	**1.000**±0.000	0.998±0.005	0.949±0.068	**1.000**±0.000
Lorenz-96 (F=40)	0.981±0.007	0.896±0.007	**0.984**±0.011	0.968±0.016	0.854±0.008	**0.976**±0.013

our geometric interpretation that causality detection benefits from appropriate metric normalization. We also developed component-wise IGGC to isolate our theoretical principles from JRNGC's implementation advantages. In Table 2, we showed that information-geometric quantities effectively detect causality. While cMLP showed stronger performance on high-dimensional VAR data, IGGC exhibited good nonlinear detection capabilities. See Appendix C for implementation details.

Our framework moreover helps to explain why different architectures succeed at causality detection. For component-wise methods (cMLP, cLSTM) [22], each model learns a separate manifold slice where causal relationships emerge naturally. For additive methods such as NAVAR [6], the architecture preserves direct information flow through component-wise networks, with gradients maintaining the path-specific structure: $\partial f_{\theta,i}(x)/\partial x_j^{t-\ell} = \partial h_{ij}^{(\ell)}(x_j^{t-\ell})/\partial x_j^{t-\ell}$. This allows IGGC to decompose proportionally to the expected squared gradient of these components. For shared-model methods (JRNGC) [26], despite shared hidden layers, the differential geometry of the input-to-output transformation enables identification of variable-specific causal paths.

Reparameterization Robustness IGGC shows advantages in stability across network parameterizations, reducing mean absolute differences by 18.9% for non-linear (cascade) systems and 30.1% for linear (VAR) systems compared to standard Jacobian approaches. Visualization analysis reveals both methods struggle with complete cascade recovery ($X_1 \to X_2 \to X_3 \to X_4 \to X_5$)—readily detecting root influences but failing to isolate direct sequential links, suggesting an important direction for future research. Yet, IGGC stabilizes causal strength measurements across reasonable parameterizations, with ensembled models achieving improved structure recovery.

Table 3: Robustness: JRNGC vs IGGC

	Non-linear		**Linear**	
Metric	No Fisher	Fisher	No Fisher	Fisher
ROC AUC	0.747	0.720	0.985	0.992
PR AUC	0.640	0.587	0.878	0.890
Mean Abs. Diff	26.207	21.259 (-18.9%)	0.554	0.387 (-30.1%)
Rank Stability	0.707	0.707	0.700	0.700
Ensemble ROC	0.733	–	1.000	–

Thus, our information-geometric framework unifies neural Granger causality methods by revealing that causal relationships emerge as directional information flow on statistical manifolds, providing theoretical foundations for neural causality methods, particularly under Gaussian assumptions, while offering geometric interpretations that illuminate why these techniques succeed empirically across broader settings.

References

[1] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10 (2):251–276, February 1998. ISSN 1530-888X. doi: 10.1162/089976698300017746. URL http://dx.doi.org/10.1162/089976698300017746.

[2] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. American Mathematical Society, April 2007. ISBN 9781470446055. doi: 10.1090/mmono/191. URL http://dx.doi.org/10.1090/mmono/191.

[3] P. O. Amblard, R. Vincent, O. J. J. Michel, and C. Richard. Kernelizing geweke's measures of granger causality. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, page 1–6. IEEE, September 2012. doi: 10.1109/mlsp.2012.6349710. URL http://dx.doi.org/10.1109/MLSP.2012.6349710.

[4] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/b22b257ad0519d4500539da3c8bcf4dd-Paper.pdf.

[5] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/ACPROF:OSO/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

[6] Bart Bussmann, Jannes Nys, and Steven Latré. Neural additive vector autoregression models for causal discovery in time series. In *Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings 24*, pages 446–460. Springer, 2021.

[7] Doris Entner and Patrik O. Hoyer. On causal discovery from time series data using fci. In Petri Myllymäki, Teemu Roos, and Tommi Jaakkola, editors, *Proceedings of the 5th European Workshop on Probabilistic Graphical Models*, pages 121–128, Finland, 2010. Helsinki Institute for Information Technology HIIT. 5th European Workshop on Probabilistic Graphical Models ; Conference date: 13-09-2010 Through 15-09-2010.

[8] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine*

Learning Research, pages 297–299. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/golowich18a.html.

[9] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424, August 1969. ISSN 0012-9682. doi: 10.2307/1912791. URL http://dx.doi.org/10.2307/1912791.

[10] Craig Hiemstra and Jonathan D. Jones. Testing for linear and nonlinear granger causality in the stock price- volume relation. *The Journal of Finance*, 49(5):1639, December 1994. ISSN 0022-1082. doi: 10.2307/2329266. URL http://dx.doi.org/10.2307/2329266.

[11] Hongming Li, Shujian Yu, and Jose Principe. Causal recurrent variational autoencoder for medical time series generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8562–8570, June 2023. ISSN 2159-5399. doi: 10.1609/aaai.v37i7.26031. URL http://dx.doi.org/10.1609/aaai.v37i7.26031.

[12] Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Berlin Heidelberg, 2005. ISBN 9783540277521. doi: 10.1007/978-3-540-27752-1. URL http://dx.doi.org/10.1007/978-3-540-27752-1.

[13] Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. Kernel method for nonlinear granger causality. *Physical Review Letters*, 100(14), April 2008. ISSN 1079-7114. doi: 10.1103/physrevlett.100.144103. URL http://dx.doi.org/10.1103/PhysRevLett.100.144103.

[14] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2408–2417, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/martens15.html.

[15] Colin McDiarmid. Surveys in combinatorics, 1989: On the method of bounded differences. In *Cambridge University Press*, 1989. URL https://api.semanticscholar.org/CorpusID:116663483.

[16] Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, January 2019. ISSN 2504-4990. doi: 10.3390/make1010019. URL http://dx.doi.org/10.3390/make1010019.

[17] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/rezende15.html.

[18] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science*

Advances, 5(11), November 2019. ISSN 2375-2548. doi: 10.1126/sciadv.aau4996. URL http://dx.doi.org/10.1126/sciadv.aau4996.

[19] Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, July 2000. ISSN 1079-7114. doi: 10.1103/physrevlett.85.461. URL http://dx.doi.org/10.1103/PhysRevLett.85.461.

[20] Anil K. Seth, Adam B. Barrett, and Lionel Barnett. Granger causality analysis in neuroscience and neuroimaging. *The Journal of Neuroscience*, 35(8):3293–3297, February 2015. ISSN 1529-2401. doi: 10.1523/jneurosci.4399-14.2015. URL http://dx.doi.org/10.1523/JNEUROSCI.4399-14.2015.

[21] Stephen M. Smith, Karla L. Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F. Beckmann, Thomas E. Nichols, Joseph D. Ramsey, and Mark W. Woolrich. Network modelling methods for fmri. *NeuroImage*, 54(2):875–891, January 2011. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2010.08.063. URL http://dx.doi.org/10.1016/j.neuroimage.2010.08.063.

[22] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2021. ISSN 1939-3539. doi: 10.1109/tpami.2021.3065601. URL http://dx.doi.org/10.1109/tpami.2021.3065601.

[23] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000. URL https://arxiv.org/abs/physics/0004057.

[24] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

[25] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, February 2019. ISBN 9781108498029. doi: 10.1017/9781108627771. URL http://dx.doi.org/10.1017/9781108627771.

[26] Wanqi Zhou, Shuanghao Bai, Shujian Yu, Qibin Zhao, and Badong Chen. Jacobian regularizer-based neural Granger causality. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 61763–61782. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhou24a.html.

Appendix A. Extended Background and Related Work

A.1. Granger Causality

Clive Granger formalized his causal framework in 1969 [9], building on Wiener's theory that if prediction of variable B improves by incorporating the past information of variable A, then A causes B. Formally, for a multivariate time series $\mathbf{x} = \{x_1, x_2, \ldots, x_D\}$ with D dimensions, the time series is modeled as

$$x_j(t) = f_j\Big(x_1(<t), \ldots, x_D(<t)\Big) + \varepsilon_j, \tag{8}$$

where ε_j is an independent noise term, and $x_i(<t)$ denotes the past information of time series x_i. The prediction of the value of x_j at time t depends on the past information of other time series, which are the potential causes or "parent" of x_j. Traditional Granger causality testing involved Vector Autoregressive (VAR) models, that is

$$x_i(t) = \sum_{j=1}^{D} \sum_{\ell=1}^{\tau} A_{ij}^{(\ell)} x_j(t-\ell) + \varepsilon_i(t), \tag{9}$$

where τ is the maximum lag and $\varepsilon_i(t)$ is noise. The coefficients $A_{ij}^{(\ell)}$ encode causal relationships, with $A_{ij}^{(\ell)} = 0$ implying no Granger causality from j to i at lag ℓ [12].

A.2. Nonlinear Extensions

Several approaches extend Granger causality to nonlinear systems. Schreiber [19] reformulated Granger causality in information-theoretic terms as

$$TE_{j\to i} = \sum p\big(x_i^{t+1}, x_i^{<t}, x_j^{<t}\big) \log \frac{p\big(x_i^{t+1} \mid x_i^{<t}, x_j^{<t}\big)}{p\big(x_i^{t+1} \mid x_i^{<t}\big)}. \tag{10}$$

Transfer entropy (TE) captures nonlinear dependencies but requires density estimation, which scales poorly with dimensionality. Kernel Granger causality [3, 13] maps time series into reproducing kernel Hilbert spaces to capture nonlinear relationships while maintaining computational tractability. However, kernel selection remains challenging. Regarding constraint-based methods, approaches like PCMCI [18] and tsFCI [7] use conditional independence tests to infer causal graphs from time series. While theoretically sound, they struggle with high dimensions and require multiple hypothesis tests.

A.3. Neural Granger Causality

Recent work leverages neural networks' representation power for Granger causality. Tank *et al.* [22] introduced neural Granger causality using multilayer perceptrons (cMLP) and LSTMs (cLSTM). They train separate models for each target variable with sparsity constraints as defined in Eq. (1). These component-wise models effectively capture complex dependencies but offer limited theoretical justification for why sparsity in first-layer weights corresponds to causal relationships. NAVAR [6] uses additive structure given by

$$f_{\theta,i}(\mathbf{x}) = \sum_{j=1}^{d} \sum_{\ell=1}^{\tau} h_{ij}^{(\ell)}\Big(x_j^{t-\ell}\Big), \tag{11}$$

where $h_{ij}^{(\ell)}$ are neural networks. This structure preserves interpretability while enabling complex nonlinear modeling. Then, several methods impose structural constraints. TCDF [16] uses attention mechanisms with dilated convolutions, implicitly assuming attention weights capture causal importance. CR-VAE [11] combines variational autoencoders with causal discovery, leveraging latent representations without rigorously establishing how these relate to causality. JRNGC [26] addresses computational limitations by using a single model with Jacobian regularization as defined in Eq. (2). The method uses Jacobian magnitude to quantify causal strength, that is

$$GC_{j\to i} = \sum_{\ell=1}^{\tau} \left| \frac{\partial f_i(\mathbf{x})}{\partial x_j^{t-\ell}} \right|^2. \tag{12}$$

JRNGC displays strong empirical performance with improved computational efficiency. However, it exemplifies the theory-practice gap in neural causality: while the Jacobian intuitively measures how changes in inputs affect outputs, there is no rigorous explanation for why this quantity specifically corresponds to Granger causality.

Information geometry provides a framework for understanding how probability distributions change—potentially offering the theoretical foundation missing in neural causality methods. However, this connection has not been previously explored, leaving a significant opportunity to bridge neural networks and statistical manifolds in the context of causal discovery.

Appendix B. Information-Geometric Framework

B.1. Neural Statistical Manifolds

Next, we provide a comprehensive derivation of how neural networks induce statistical manifolds and how information geometry provides a framework for understanding causality. For a neural network $f_\theta : \mathbb{R}^{d\tau} \to \mathbb{R}^d$ with parameters $\theta \in \Theta \subseteq \mathbb{R}^k$, we generally assume a conditional probability model as

$$p_\theta(y \mid x) = \prod_{i=1}^{d} p_\theta(y_i \mid x). \tag{13}$$

For regression tasks with Gaussian noise, this takes the form

$$p_\theta(y \mid x) = \mathcal{N}\big(f_\theta(x), \Sigma_\theta\big), \tag{14}$$

where Σ_θ is the covariance matrix (often assumed diagonal for simplicity). The statistical manifold $\mathcal{M} = \big\{p_\theta(\cdot \mid x) : \theta \in \Theta\big\}$ can be equipped with the Fisher Information Metric in parameter space $G_{ij}(\theta)$ defined in Eq. (3). However, for causal analysis, we need to measure how perturbations in input variables affect the output distribution. This requires transforming the Fisher metric from parameter space to input space (Definition 2). This input-space metric naturally quantifies how information flows through the statistical manifold as inputs change, providing the geometric foundation for measuring causal relationships.

B.2. Causal Information Flow (Proof of Proposition 1)

Proof For a neural network $f_\theta : \mathbb{R}^{d\tau} \to \mathbb{R}^d$ with parameters $\theta \in \Theta \subseteq \mathbb{R}^k$, the gradient with respect to input is given by

$$v_{j\to i} = \nabla_{x_j} \log p_\theta(y_i \mid x) = \frac{\partial \log p_\theta(y_i \mid x)}{\partial x_j}. \tag{15}$$

By the chain rule through the neural network output, we obtain

$$\frac{\partial \log p_\theta(y_i \mid x)}{\partial x_j} = \frac{\partial \log p_\theta(y_i \mid x)}{\partial f_{\theta,i}(x)} \frac{\partial f_{\theta,i}(x)}{\partial x_j}. \tag{16}$$

For common exponential family distributions, this takes specific forms. For example, for a Gaussian distribution with mean $f_{\theta,i}(x)$ and variance σ^2, we have

$$\frac{\partial \log p_\theta(y_i \mid x)}{\partial x_j} = \frac{1}{\sigma^2}\Big(y_i - f_{\theta,i}(x)\Big) \frac{\partial f_{\theta,i}(x)}{\partial x_j}. \tag{17}$$

Following Definition 2, we define the Fisher Information Matrix directly in input space for the conditional distribution $p_\theta(y|x)$ as

$$G_x(x) = \mathbb{E}_{y|x}\Big[\Big(\nabla_x \log p_\theta(y \mid x)\Big)\Big(\nabla_x \log p_\theta(y \mid x)\Big)^T\Big], \tag{18}$$

where ∇_x denotes the gradient with respect to input x. This provides the natural Riemannian metric for measuring information flow in the input space. The magnitude of causal flow is thus

$$\|v_{j\to i}\|^2_{G_x} = v_{j\to i}^T\, G_x(x)\, v_{j\to i}. \tag{19}$$

This represents information flow through the statistical manifold from x_j to y_i. The interpretation follows from [2]'s work on information geometry, where gradients on statistical manifolds represent directions of maximum information change [1]. ■

B.3. Full Definition and Properties of IGGC

From Definition 3, we write the Information-Geometric Granger Causality (IGGC) from variable j to variable i at lag ℓ as

$$\text{IGGC}^{(\ell)}_{j\to i} = \mathbb{E}_{x\sim p(x)}\left[\left\|\nabla_{x_j^{t-\ell}} \log p_\theta\Big(x_i^{t+1} \mid x^{t-\tau:t}\Big)\right\|^2_{G_x(x)}\right], \tag{5}$$

where $\|\cdot\|_{G_x(x)}$ is the norm induced by the input-space Fisher metric $\|v\|^2_{G_x} = v^T G_x(x) v$. This formulation provides several key advantages. It naturally extends to multivariate settings, and handles non-stationarity through time-indexed gradients. It also accounts for model uncertainty through the Fisher metric, and provides a unified framework for diverse neural architectures. For full-time Granger causality, we obtain a tensor of IGGC values across all variable pairs and lags. For summary Granger causality, we aggregate across lags as

$$\text{IGGC}_{j\to i} = \sum_{\ell=1}^{\tau} \text{IGGC}^{(\ell)}_{j\to i}. \tag{20}$$

B.4. Fundamental Properties of IGGC (Proof of Theorem 1)

Proof We prove each property in turn.

Part (1): Non-negativity Using Definition 3, and since the input-space Fisher metric $G_x(x)$ is positive semi-definite, we have for any vector v:

$$\|v\|_{G_x}^2 = v^T G_x(x) v \geq 0. \tag{21}$$

The expectation of a non-negative quantity is non-negative, hence $\text{IGGC}_{j\to i}^{(\ell)} \geq 0$.

Part (2): Causal identification
($\Rightarrow$) Suppose $\text{IGGC}_{j\to i}^{(\ell)} = 0$. Then, we have

$$\mathbb{E}_{x\sim p(x)}\left[\left\|\nabla_{x_j^{t-\ell}} \log p_\theta(x_i^{t+1} \mid x^{t-\tau:t})\right\|_{G_x}^2\right] = 0 \tag{22}$$

Since the integrand is non-negative and the expectation equals zero, we must have

$$\nabla_{x_j^{t-\ell}} \log p_\theta\left(x_i^{t+1} \mid x^{t-\tau:t}\right) = 0 \quad \text{almost surely.} \tag{23}$$

This implies that $p_\theta\left(x_i^{t+1} \mid x^{t-\tau:t}\right)$ does not depend on $x_j^{t-\ell}$, that is

$$p_\theta\left(x_i^{t+1} \mid x^{t-\tau:t}\right) = p_\theta\left(x_i^{t+1} \mid x_{-j}^{t-\tau:t}\right). \tag{24}$$

This is precisely the definition of conditional independence $X_j^{t-\ell} \perp\!\!\!\perp X_i^{t+1} \mid X_{-j}^{t-\tau:t}$.
($\Leftarrow$) Conversely, if $X_j^{t-\ell} \perp\!\!\!\perp X_i^{t+1} \mid X_{-j}^{t-\tau:t}$, then $p_\theta\left(x_i^{t+1} \mid x^{t-\tau:t}\right)$ does not depend on $x_j^{t-\ell}$, so we have

$$\frac{\partial \log p_\theta\left(x_i^{t+1} \mid x^{t-\tau:t}\right)}{\partial x_j^{t-\ell}} = 0. \tag{25}$$

Therefore, $\text{IGGC}_{j\to i}^{(\ell)} = 0$.

Part (3): Scale consistency
Under rescaling of the noise variance $\sigma^2 \to c\,\sigma^2$, by Theorem 2, we have $\text{IGGC}_{j\to i}^{(\ell)} \to c^{-1}\,\text{IGGC}_{j\to i}^{(\ell)}$. This scaling ensures that causal relationships remain detectable across different noise levels, with appropriately adjusted strength measures that reflect the signal-to-noise ratio. ■

B.5. JRNGC as Approximation of IGGC (Proof of Theorem 2)

Proof We prove the equivalence between IGGC and JRNGC under Gaussian assumptions with explicit treatment of the Fisher metric.

Setup: Consider the Gaussian conditional distribution $p_\theta(y_i \mid x) = \mathcal{N}(f_{\theta,i}(x), \sigma^2)$ where $f_{\theta,i}(x)$ is the i-th output of neural network f_θ.

We first compute the log-likelihood partial derivative, that is

$$\log p_\theta(y_i \mid x) = -\frac{1}{2}\log\Big(2\pi\sigma^2\Big) - \frac{1}{2\sigma^2}\Big(y_i - f_{\theta,i}(x)\Big)^2 \tag{26}$$

Taking the partial derivative with respect to $x_j^{t-\ell}$, we have

$$\frac{\partial \log p_\theta(y_i \mid x)}{\partial x_j^{t-\ell}} = \frac{1}{\sigma^2}\Big(y_i - f_{\theta,i}(x)\Big)\frac{\partial f_{\theta,i}(x)}{\partial x_j^{t-\ell}} \tag{27}$$

For the scalar partial derivative, the corresponding Fisher Information Matrix element is given by

$$\Big[G_x(x)\Big]_{j,\ell} = \mathbb{E}_{y_i|x}\left[\left(\frac{\partial \log p_\theta(y_i \mid x)}{\partial x_j^{t-\ell}}\right)^2\right] \tag{28}$$

$$= \mathbb{E}_{y_i|x}\left[\frac{1}{\sigma^4}\Big(y_i - f_{\theta,i}(x)\Big)^2\left(\frac{\partial f_{\theta,i}(x)}{\partial x_j^{t-\ell}}\right)^2\right] \tag{29}$$

$$= \frac{1}{\sigma^4}\mathbb{E}_{y_i|x}\Big[\big(y_i - f_{\theta,i}(x)\big)^2\Big]\left(\frac{\partial f_{\theta,i}(x)}{\partial x_j^{t-\ell}}\right)^2. \tag{30}$$

Since $y_i \mid x \sim \mathcal{N}(f_{\theta,i}(x), \sigma^2)$, we have $\mathbb{E}_{y_i|x}\Big[(y_i - f_{\theta,i}(x))^2\Big] = \sigma^2$, therefore

$$\Big[G_x(x)\Big]_{j,\ell} = \frac{1}{\sigma^2}\left(\frac{\partial f_{\theta,i}(x)}{\partial x_j^{t-\ell}}\right)^2. \tag{31}$$

For a scalar partial derivative, the Fisher norm squared is

$$\left\|\frac{\partial \log p_\theta\Big(x_i^{t+1} \mid x^{t-\tau:t}\Big)}{\partial x_j^{t-\ell}}\right\|^2_{G_x(x)} = \left(\frac{\partial \log p_\theta\Big(x_i^{t+1} \mid x^{t-\tau:t}\Big)}{\partial x_j^{t-\ell}}\right)^2\Big[G_x(x)\Big]_{j,\ell}. \tag{32}$$

At prediction time, we substitute $y_i = x_i^{t+1}$ and use the fact that under the learned model, we have

$$\mathbb{E}_{x_i^{t+1}|x^{t-\tau:t}}\Big[\big(x_i^{t+1} - f_{\theta,i}(x^{t-\tau:t})\big)^2\Big] = \sigma^2. \tag{33}$$

Therefore,

$$\begin{aligned}
\text{IGGC}_{j\to i}^{(\ell)} &= \mathbb{E}_{x\sim p(x)}\left[\left\|\frac{\partial \log p_\theta(x_i^{t+1} \mid x^{t-\tau:t})}{\partial x_j^{t-\ell}}\right\|^2_{G_x(x)}\right] \\
&= \mathbb{E}_{x\sim p(x)}\left[\frac{1}{\sigma^4}\Big(x_i^{t+1} - f_{\theta,i}(x^{t-\tau:t})\Big)^2\left(\frac{\partial f_{\theta,i}(x^{t-\tau:t})}{\partial x_j^{t-\ell}}\right)^2\frac{1}{\sigma^2}\left(\frac{\partial f_{\theta,i}(x^{t-\tau:t})}{\partial x_j^{t-\ell}}\right)^2\right] \\
&= \frac{1}{\sigma^2}\mathbb{E}_{x\sim p(x)}\left[\left(\frac{\partial f_{\theta,i}(x^{t-\tau:t})}{\partial x_j^{t-\ell}}\right)^2\right],
\end{aligned} \tag{34}$$

where the last step uses $\mathbb{E}\Big[(x_i^{t+1} - f_{\theta,i}(x^{t-\tau:t}))^2\Big] = \sigma^2$. We know that the JRNGC Jacobian regularization term for the $(j,\ell) \to i$ component is exactly given by

$$\mathbb{E}\left[\left(\frac{\partial f_{\theta,i}(x)}{\partial x_j^{t-\ell}}\right)^2\right]. \tag{35}$$

Thus, we have

$$\text{IGGC}_{j\to i}^{(\ell)} = \frac{1}{\sigma^2}\text{JRNGC}_{j\to i}^{(\ell)}. \tag{36}$$

This establishes that JRNGC's Jacobian regularization is exactly the IGGC measure scaled by σ^{-2} under Gaussian assumptions.

Remark 2 *The proof assumes that the Fisher metric can be treated component-wise for scalar partial derivatives, which corresponds to the diagonal Fisher approximation commonly used in practice. The exact equivalence holds under this assumption, which is reasonable when inputs have limited cross-correlations or when computational efficiency requires diagonal approximations.*

■

Remark 3 *While Theorem 2 establishes exact equivalence under Gaussian assumptions, the broader information-geometric interpretation of IGGC as measuring directional information flow provides qualitative insights for understanding neural causality methods even when exact mathematical equivalence does not hold.*

B.6. Consistency of Regularized Estimator (Proof of Theorem 5)

Setup Assume the loss function ℓ is L-Lipschitz, the parameter space Θ is compact with diameter D, and the population risk $L(\theta) = \mathbb{E}\big[\ell\big(f_\theta(X), Y\big)\big]$ is m-strongly convex near the minimizer θ_0.

Theorem 5 *Under these conditions, the regularized estimator satisfies*

$$\left\|\hat{\theta}_\lambda - \theta^*\right\| = O_p\left(\sqrt{\frac{k\log d}{n}} + \lambda\right). \tag{37}$$

Proof Define $L_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(f_\theta(x_i), y_i)$ the empirical risk, $L(\theta) = \mathbb{E}[\ell(f_\theta(X), Y)]$ the population risk, and $R(\theta)$ the regularizer. The regularized estimator is

$$\hat{\theta}_\lambda = \arg\min_{\theta\in\Theta}\Big\{L_n(\theta) + \lambda R(\theta)\Big\}. \tag{38}$$

By decomposition, we have

$$\left\|\hat{\theta}_\lambda - \theta^*\right\| \le \left\|\hat{\theta}_\lambda - \theta_0\right\| + \|\theta_0 - \theta^*\|, \tag{39}$$

where $\theta_0 = \arg\min_\theta L(\theta)$. Then, by McDiarmid's inequality [15], for Lipschitz losses, we obtain

$$P\left(\sup_{\theta\in\Theta}|L_n(\theta) - L(\theta)| > t\right) \le 2\exp\left(-\frac{2nt^2}{L^2}\right). \tag{40}$$

Setting $t = L\sqrt{\frac{\log(2d/\delta)}{2n}}$, with probability $1-\delta$, we have

$$\sup_{\theta\in\Theta} |L_n(\theta) - L(\theta)| \leq L\sqrt{\frac{\log(2d/\delta)}{2n}}. \tag{41}$$

By optimality of $\hat{\theta}_\lambda$, we write

$$L_n(\hat{\theta}_\lambda) + \lambda R(\hat{\theta}_\lambda) \leq L_n(\theta_0) + \lambda R(\theta_0). \tag{42}$$

Using strong convexity and the concentration result in Eq. (41) [5], we obtain

$$\frac{m}{2}\left\|\hat{\theta}_\lambda - \theta_0\right\|^2 \leq 2L\sqrt{\frac{\log(2d/\delta)}{2n}} + \lambda\Big(R(\theta_0) - R(\hat{\theta}_\lambda)\Big). \tag{43}$$

If R is bounded by B, we have

$$\left\|\hat{\theta}_\lambda - \theta_0\right\| \leq \sqrt{\frac{8L}{m}}\sqrt{\frac{\log(2d/\delta)}{2n}} + \frac{2\lambda B}{m}. \tag{44}$$

With appropriate constants, the final bound is given by

$$\left\|\hat{\theta}_\lambda - \theta^*\right\| = O_p\left(\sqrt{\frac{k\log d}{n}} + \lambda\right). \tag{45}$$

The factor k appears from covering number arguments for k-dimensional parameter spaces. ■

B.7. Asymptotic Consistency (Proof of Theorem 3)

Setup Assume identifiability: for the true parameter θ^*, if $\text{IGGC}^{(\ell)}_{j\to i}[\theta^*] = 0$, then the true causal graph has no edge from j to i at lag ℓ.

Proof From Theorem 5 and continuity of IGGC in θ (follows from smoothness of neural networks), we have

$$\sup_{i,j,\ell}\left|\widehat{\text{IGGC}}^{(\ell)}_{j\to i} - \text{IGGC}^{(\ell)}_{j\to i}[\theta^*]\right| = O_p\left(\sqrt{\frac{k\log d}{n}} + \lambda_n\right). \tag{46}$$

Then, define the minimum gap as

$$\Delta = \min\Big\{\text{IGGC}^{(\ell)}_{j\to i}[\theta^*] : (i,j,\ell) \in G^*\Big\}. \tag{47}$$

Under identifiability, $\Delta > 0$.

Using threshold $\eta_n = \Delta/2$, we recover edge (i,j,ℓ) if $\widehat{\text{IGGC}}^{(\ell)}_{j\to i} > \eta_n$. For correct recovery, we have

$$\sup_{i,j,\ell}\left|\widehat{\text{IGGC}}^{(\ell)}_{j\to i} - \text{IGGC}^{(\ell)}_{j\to i}[\theta^*]\right| < \frac{\Delta}{2}. \tag{48}$$

This occurs with probability approaching 1 when

$$\sqrt{\frac{k \log d}{n}} + \lambda_n < \frac{\Delta}{2}. \tag{49}$$

Given $\lambda_n \to 0$ and $\lambda_n \sqrt{n} \to \infty$, the bias term $\lambda_n \to 0$, the variance term $\sqrt{k \log d/n} \to 0$, and the condition $\lambda_n \sqrt{n} \to \infty$ ensures sufficient regularization. Therefore, we have

$$P\Big(\hat{G}_n = G^*\Big) \geq P\left(\sup_{i,j,\ell} \left|\widehat{\text{IGGC}}^{(\ell)}_{j\to i} - \text{IGGC}^{(\ell)}_{j\to i}[\theta^*]\right| < \frac{\Delta}{2}\right) \to 1 \quad \text{as } n \to \infty. \tag{50}$$

■

B.8. Finite-Sample Complexity (Proof of Theorem 4)

Setup Assume neural networks with bounded weights $\|\theta\|_\infty \leq B$, Lipschitz activation functions, and minimum detectable effect size ε.

Proof We define the function class as

$$\mathcal{F} = \left\{ x \mapsto \frac{\partial f_{\theta,i}(x)}{\partial x_j^{t-\ell}} : \|\theta\|_\infty \leq B, i, j \in [d], \ell \in [\tau] \right\}. \tag{51}$$

For L-layer neural networks with bounded weights, the Rademacher complexity [4, 8] is given by

$$\mathcal{R}_n(\mathcal{F}) \leq O\left(\frac{B^L L^{3/2}}{\sqrt{n}} \sqrt{k \log k}\right). \tag{52}$$

For the squared functions in IGGC, using Rademacher complexity bounds, we obtain

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f^2(x_i) - \mathbb{E}\Big[f^2(X)\Big] \right|\right] \leq 2B^2 \mathcal{R}_n(\mathcal{F}). \tag{53}$$

With probability $1 - \delta/2$,

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f^2(x_i) - \mathbb{E}\Big[f^2(X)\Big] \right| \leq 2B^2 \mathcal{R}_n(\mathcal{F}) + B^2 \sqrt{\frac{2 \log(2/\delta)}{n}}. \tag{54}$$

For accurate IGGC estimation with error at most $\varepsilon/2$, we have

$$2B^2 \mathcal{R}_n(\mathcal{F}) + B^2 \sqrt{\frac{2 \log(2/\delta)}{n}} \leq \frac{\varepsilon}{2}. \tag{55}$$

Substituting the Rademacher bound, we obtain

$$n \geq O\left(\frac{B^{2L+4} L^3 k \log k}{\varepsilon^2} + \frac{B^4 \log(2/\delta)}{\varepsilon^2}\right). \tag{56}$$

With $d^2\tau$ potential edges, using union bound, we have

$$n \geq O\left(\frac{B^{2L+4}L^3 k \log k}{\varepsilon^2} + \frac{B^4 \log(2d^2\tau/\delta)}{\varepsilon^2}\right). \tag{57}$$

From standard concentration inequalities for parameter estimation [25], we write

$$n \geq O\left(\frac{k \log(k/\delta)}{\varepsilon^2}\right). \tag{58}$$

Combining both requirements and simplifying constants give

$$n = O\left(\frac{d^2\tau \log(d/\delta)}{\varepsilon^2} + \frac{k \log(k/\delta)}{\varepsilon^2}\right). \tag{59}$$

The first term handles causal discovery complexity, the second handles parameter estimation [24]. ■

B.9. Extensions to Other Neural Causality Methods

In this section, we further explain that our information-geometric framework applies beyond JRNGC to explain diverse neural approaches to Granger causality. The key insight is that any neural network learning conditional distributions induces a statistical manifold where causality emerges geometrically. Component-wise methods such as cMLP and cLSTM [22] train separate models f_{θ_i} for each target variable i. In our framework, each model learns a manifold slice. The separate IGGC values, defined in Eq. (5), naturally emerge from these individual manifolds. This explains why component-wise training captures causality despite computational inefficiency.

On the other hand, NAVAR [6] uses additive structure as

$$f_{\theta,i}(x) = \sum_{j=1}^{d} \sum_{\ell=1}^{\tau} h_{ij}^{(\ell)}\left(x_j^{t-\ell}\right), \tag{60}$$

where $h_{ij}^{(\ell)}$ are neural networks. The gradient with respect to input decomposes according to this structure as

$$\frac{\partial f_{\theta,i}(x)}{\partial x_j^{t-\ell}} = \frac{\partial h_{ij}^{(\ell)}\left(x_j^{t-\ell}\right)}{\partial x_j^{t-\ell}}. \tag{61}$$

Therefore, by Theorem 2, IGGC decomposes as

$$\text{IGGC}_{j\to i}^{(\ell)} = \frac{1}{\sigma^2}\mathbb{E}\left[\left(\frac{\partial h_{ij}^{(\ell)}(x_j^{t-\ell})}{\partial x_j^{t-\ell}}\right)^2\right]. \tag{62}$$

This preserves the geometric interpretation: each component $h_{ij}^{(\ell)}$ contributes directly to the information flow from j to i at lag ℓ.

Then, methods such as TCDF [16] use attention mechanisms to compute weights $\alpha_{ij}^{(\ell)}$ that determine the importance of input $x_j^{t-\ell}$ for predicting x_i^{t+1}. While these attention weights are not explicitly designed to capture causal information flow, we can establish a relationship to our framework.

Proposition 2 *Under certain conditions on the attention mechanism, the attention weights can be related to the magnitude of causal information flow.*

Briefly, a standard self-attention mechanism computes weights as

$$\alpha_{ij}^{(\ell)} = \text{softmax}\left(\frac{Q_i K_j^{t-\ell}}{\sqrt{d_k}}\right), \tag{63}$$

where Q_i and $K_j^{t-\ell}$ are query and key vectors. For prediction tasks, these weights determine how much input $x_j^{t-\ell}$ contributes to the output $\hat{x}_i^{t+1}$. By the chain rule, we have

$$\frac{\partial \hat{x}_i^{t+1}}{\partial x_j^{t-\ell}} = \frac{\partial \hat{x}_i^{t+1}}{\partial \alpha_{ij}^{(\ell)}} \frac{\partial \alpha_{ij}^{(\ell)}}{\partial x_j^{t-\ell}} + \text{other terms}. \tag{64}$$

Under Gaussian assumptions, IGGC is proportional to the expected squared Jacobian (Theorem 2). When attention mechanisms are trained to minimize prediction error, they implicitly learn to assign higher weights to inputs that reduce uncertainty (i.e., that have high information content). This creates an implicit relationship where attention weights tend to be correlated with magnitudes of causal influence. The precise relationship depends on the specific attention architecture, but empirically, we observe

$$\text{corr}\left(\alpha_{ij}^{(\ell)}, \left\|\nabla_{x_j^{t-\ell}} \log p_\theta(x_i^{t+1} \mid x^{t-\tau:t})\right\|\right) > 0. \tag{65}$$

This connection provides a theoretical justification for why attention-based methods can successfully discover causal relationships, even though they were not explicitly designed with information geometry.

B.10. Limitations and Scope

While our framework provides exact theoretical foundations under Gaussian assumptions (Theorem 2), real-world applications often involve non-Gaussian time series. In such cases, our IGGC measure serves as a principled approximation that preserves the geometric intuition of measuring directional information flow, even when mathematical equivalence is not guaranteed. Empirical validation on diverse datasets suggests the framework's practical utility extends beyond its theoretical guarantees.

Appendix C. Implementation Details

C.1. Neural Network Architectures

For the experiments, we implemented our IG-NGC framework using a residual MLP architecture similar to JRNGC [26], but with explicit accounting for Fisher metric normalization:

$$\hat{\mathbf{x}}_{1:D}^t = \text{FC}_2\Big(\text{ResidualBlock}\big(\text{FC}_1(\mathbf{x}_{1:D}^{t-\tau:t-1})\big)\Big). \tag{66}$$

The residual blocks follow the structure:

$$\text{ResidualBlock}(h) = \text{LayerNorm}\big(h + g(h)\big) \tag{67}$$

$$g(h) = \text{Dropout}\Big[\text{ReLU}\Big[\text{FC}_{\text{hidden}}\Big(\text{WeightNorm}\big(\text{FC}_{\text{input}}(h)\big)\Big)\Big]\Big]. \tag{68}$$

For component-wise IGGC experiments, we used separate networks for each target variable following the cMLP structure from [22] but with added Fisher normalization.

C.2. Fisher Information Approximation

Computing the full Fisher Information Matrix would require $O(k^2)$ space for k parameters. We used the empirical Fisher approximation [14]:

$$\hat{G}_{\text{diag}}(\theta) = \text{diag}\left(\frac{1}{N}\sum_{n=1}^{N} g_n \odot g_n\right), \tag{69}$$

where g_n is the gradient of the log-likelihood for the n-th sample. This diagonal approximation maintains the essential geometric structure while ensuring scalability.

C.3. Efficient Jacobian Computation

We implemented efficient batched Jacobian computation following [26], as shown in Algorithm 1.

Algorithm 1 Efficient Jacobian Computation via Batched Autodiff

Input: Network f_θ, batch $\mathbf{X} \in \mathbb{R}^{B\times d\times\tau}$
Output: Jacobian $\mathbf{J} \in \mathbb{R}^{B\times d\times d\times\tau}$
Flatten inputs: $\mathbf{X}' = \text{reshape}(\mathbf{X}, [B, d\tau])$
Compute outputs: $\mathbf{Y} = f_\theta(\mathbf{X}')$
Initialize $\mathbf{J} = \text{zeros}([B, d, d, \tau])$
for $i = 1$ **to** d **do**
  $\text{grad}_i = \text{autograd}(\mathbf{Y}[:, i], \mathbf{X}')$; `// Parallelize over output dimensions`
  $\mathbf{J}[:, i, :, :] = \text{reshape}(\text{grad}_i, [B, d, \tau])$
end
return J

C.4. Dataset Descriptions

VAR model: We simulated VAR processes with dimensions $D \in \{10, 50, 100\}$, maximum lag $\tau \in \{3, 5\}$, and maximum estimated lag $\eta \in \{5, 10\}$. The coefficients were generated to ensure stationarity, with sparsity level 0.2 (20% of possible connections are non-zero).

Lorenz-96: We simulated the Lorenz-96 system with $D = 10$ dimensions and forcing constants $F \in \{10, 40\}$, where higher F values produce more chaotic behavior. Time series of length 500 were generated with observation noise $\sigma = 0.1$.

fMRI: We used the Smith et al. [21] simulated fMRI dataset, focusing on subject 1 from the third simulation set. This dataset has 15 regions with known ground-truth connectivity and 200 time points.