IDENTIFIABILITY GUARANTEES FOR DRIVERS AND DYNAMICS OF DELAYED PHYSICAL SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning the dynamics of a physical system directly from observations is a key problem in natural sciences, where physical consistency and interpretability are essential. A wide range of methods have been proposed, including physics-informed neural networks, which are powerful but do not guarantee identifiability of the dynamics, symbolic regression, which requires a set of precomputed operations, and causal discovery, which is more principled but usually relies on strong assumptions that physical systems may violate. In this work, we develop a theory-grounded method and prove that under a set of permissive assumptions, the structural drivers and drift of stochastic delayed differential equations are identifiable. Our method outperforms others on a benchmark for driver identifiability, and on a second benchmark to evaluate physical consistency of the learned dynamics.

1 INTRODUCTION

The evolution of many physical systems depends not only on their current state but also on their past states. For example, the climate system and ecological or neural population dynamics are driven by delayed feedbacks between variables (Chekroun & Liu, 2024; Keane et al., 2017; Çetin et al., 2026). Lagged dynamical systems explicitly model both instantaneous and time-delayed dependencies, which improves the predictive capacity and interpretability of the learned model. Such systems are formally described by stochastic delayed differential equations (SDDEs) (Mohammed, 1982; Scheutzow, 1984) which allow for delayed connections, as opposed to stochastic differential equations (SDEs) used to describe Markovian dynamical systems (Itô, 1944; 1951).

Lagged connections in physical systems such as the climate system typically lead to oscillations and chaotic behavior, characterized by ergodicity and chaoticity (Ghil et al., 2008; Ghil & Lucarini, 2020). Ergodic means that long-time statistics are identical across almost all initial conditions (Walters, 1999), while chaotic means that an infinitesimally small difference between two initial conditions will lead to two paths diverging exponentially fast (Dorfman, 1999).

Scientific machine learning (ML) has mostly focused on modeling SDE-driven systems. Neural ordinary differential equations (ODEs) (Chen et al., 2018; Dupont et al., 2019) allow the direct approximation of the gradient of a physical system from irregular sparse observations (Zhang et al., 2020). They have been extended to capture delay (Eggen & Midtfjord, 2023; Oh et al., 2025), but still do not learn the long-term statistics of the system, leading to unrealistic long-horizon trajectories. Indeed, Park et al. (2024) have shown that data-driven neural parameterizations of chaotic ordinary or partial differential equations (PDEs) fail to generalize when modeling ergodic chaotic systems, despite low test error. Physics-informed neural networks (Yin et al., 2021), or symbolic regression methods (Udrescu et al., 2020; Brunton et al., 2016) improve generalization but require a known parametric form or a set of precomputed functions and are computationally expensive.

Motivated by the above limitations, we aim to learn the true drivers and drift of an SDDE. They must therefore be identifiable, i.e. uniquely recoverable from observations. Causal discovery methods have emerged as a promising avenue to infer cause–effect relationships from time-series observations and have been extended to complex, nonlinear, lagged causal structures (Runge, 2020; Yao et al., 2022; Lippe et al., 2022; Hickman et al., 2025; Brouillard et al., 2026). To identify the causal graph, these methods typically rely on strong constraints such as acyclicity (Pamfil et al., 2020) or no instantaneous connections, which physical systems often violate. Another approach is to embed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

traditional regularization and feature selection methods into Neural ODEs (NODEs, Aliee et al., 2025), with identifiability guarantees derived for Markovian systems (Bellot et al., 2022).

In this work, we provide identifiability guarantees for ergodic, chaotic SDDE-driven systems, both at population level and in finite-sample settings. We use L_0 regularization (Louizos et al., 2018), which has been proposed in the disentanglement literature (Lachapelle et al., 2023) and to regularize neural ODEs (Aliee et al., 2022). We show that under a set of permissive assumptions, it is sufficient for identifying the correct lagged causal connections and drift of SDDEs.

Main contributions:

- We develop a theory-grounded method (L_0-Neural DDE, L_0-NDDE) and provide novel identifiability proofs for lagged and instantaneous drivers and drift of SDDE systems.
- We demonstrate that L_0-NDDE outperforms other data-driven methods for identifying structural drivers of dynamical systems on a comprehensive benchmark.
- We show that L_0-NDDE accurately captures long-term statistics on 40 simulated low-dimensional chaotic systems from the Dysts dataset (Gilpin, 2021).

2 Related work

Several causal discovery approaches have been proposed for time-series data. Building upon Granger causality (Granger, 1969), which fails to capture nonlinear dynamics (Sugihara et al., 2012), methods such as Neural Granger Causality (NGC) (Tank et al., 2022), CUTS+ (Cheng et al., 2024) or Temporal Causal Discovery Framework (TCDF) (Nauta et al., 2019) use various deep-learning techniques to learn nonlinear autoregressive links and identify lagged causal links directly from multivariate sequences. Constraint-based methods such as PCMCI+ (Runge, 2020) or F-PCMCI (Castri et al., 2023) infer causal structure using conditional independence tests between lagged variables, while permutation-based methods (GRaSP, Lam et al., 2022) search over permutations of variables and iteratively prune spurious edges to identify sparse causal graphs. Score-based methods such as DYNOTEARS (Pamfil et al., 2020) infer causal relationships by optimizing a score function with differentiable constraints that ensure identifiability of the causal graph (Zheng et al., 2018; Brouillard et al., 2026; Hickman et al., 2025). Noise-based methods (VARLiNGAM, Hyvärinen et al., 2010; RCD, Maeda & Shimizu, 2020) leverage the statistical independence of the noise term with respect to the inputs to identify causal connections. Finally, cross-mapping techniques discover lagged causal relationships by aligning lagged time-series (TSCI, Butler et al., 2024).

These causal methods typically discretize time and rely on strong assumptions about the causal graph (e.g., acyclicity) which often violate the nature of the continuous-time dynamics (Tagliapietra et al., 2026). To overcome this issue, a growing body of work has proposed to integrate graphical modelling approaches into Neural ODEs to identify the Jacobian of the physical system. Penalized regression has been used to estimate parameters of differential equations, with consistency guarantees (Ramsay et al., 2007; Raissi et al., 2017; Wenk et al., 2020). More recently, Bellot et al. (2022) derived identifiability results for drivers of SDEs and proposed to use adaptive group lasso (AGL, Zou, 2006) to select input features and identify drivers. This work has been applied to SDDEs without identifiability guarantees (Wu et al., 2024).

In our work, we propose to use L_0-induced sparsity (Louizos et al., 2018) on the input features to identify the instantaneous and lagged drivers of SDDEs. We derive novel theoretical guarantees for the identifiability of the dynamics of SDDEs, under a discrete, finite number of observations. L_0-induced sparsity has been used in causal representation learning (Lachapelle et al., 2023), and for regularizing neural ODEs (PathReg, Aliee et al., 2022).

3 Method

We consider the following continuous-time stochastic model, with delays $0 < \tau_1 < \cdots < \tau_q$,

$$dX_t = G(X_t,\, X_{t-\tau_1}, \ldots, X_{t-\tau_q})\, dt + \Sigma\, dW_t,$$
$$G_j(x_0, x_1, \ldots, x_q) \;=\; f_j\big(m_j^0 \odot x_0,\; m_j^1 \odot x_1, \ldots,\; m_j^q \odot x_q\big)\,, \quad M \in \{0,1\}^{D\times D\times(q+1)} \tag{1}$$

where W is a standard D-dimensional Brownian motion and $\Sigma \in \mathbb{R}^{D\times D}$ is the diffusion matrix ($\Sigma\Sigma^\top$ being the noise covariance per unit time). M is a sparse binary tensor representing the dynamical drivers at timesteps $0, \tau_1, \ldots, \tau_q$. We note $Z_t := (X_t, X_{t-\tau_1}, \ldots, X_{t-\tau_q})$ to simplify the notation when needed. We assume here that there is a maximum time lag $\tau_{\max}$ for the lagged interactions, which happen within the timesteps $0, \tau_1, \ldots, \tau_q$. f_j is the nonlinear function giving the drift (the infinitesimal conditional mean rate of change) of $X_{t,j}$ as a function of its drivers. We learn the functions $\widetilde{f}_j(\tilde{m}_j^0 \odot x_0,\ \tilde{m}_j^1 \odot x_1, \ldots,\ \tilde{m}_j^q \odot x_q) =: \widetilde{G}_j(x_0, \ldots, x_q)$ to approximate this system.

In this section, we first show that the drivers and the dynamics, i.e. the matrix M and function G, can in principle be recovered from a finite number of observations. We first give assumptions under which the SDDE (1) admits a unique strong solution, which is a prerequisite to identifiability. We then show that G and M are identifiable at the population level, either under a hard sparsity constraint (Theorem 1) or with an L_0 penalty (Theorem 2). Under additional assumptions, we prove that the system is geometrically ergodic (Theorem 3) and provide a finite-sample identifiability theorem (Theorem 4). Proofs are given in Section A and we provide a proof map (Figure 1) to help the reader understand the construction of the proofs.

3.1 Existence and uniqueness of a solution

Because the system is not Markovia

n, the evolution of the system depends on the whole history of length $\tau_{\max}$ rather than the current value only. This is represented by the segment space and the segment process (Mao, 2007; Hairer et al., 2011):

$$\mathcal{C} \;:=\; C\big([-\tau_{\max}, 0];\, \mathbb{R}^D\big), \qquad X_t^{\mathrm{seg}}(s) \;:=\; X_{t+s}, \quad s \in [-\tau_{\max}, 0],$$

where $\mathcal{C}$ is the space of continuous functions from $[-\tau_{\max}, 0]$ to $\mathbb{R}^D$, equipped with the supremum norm $\|\varphi\|_{\mathcal{C}} = \sup_{s\in[-\tau_{\max},0]}\|\varphi(s)\|$ for any given path $\varphi \in \mathcal{C}$. The segment process $(X_t^{\mathrm{seg}})_{t\geq 0}$ is then a Markov process on the (infinite-dimensional) space $\mathcal{C}$. Accordingly, the initial condition is a $\mathcal{C}$-valued random variable X_0^{seg}, i.e. all values observed in the previous time interval $[-\tau_{\max}, 0]$.

We impose standard assumptions on the underlying process: we require that the observed history does not anticipate future noise, that the noise acts in all dimensions (non-degenerate), and that the drift maintains finite energy along the trajectory, i.e. the system has finite energy.

Assumption 1 (Data-Generating Process)**.** *The true process X solves the SDDE Equation 1 on $\mathbb{R}^D$:*

(i) ***Initial conditions:*** *The initial state X_0^{seg} is independent of W, and $\mathbb{E}\|X_0^{\mathrm{seg}}\|_{\mathcal{C}}^2 < \infty$.*
(ii) ***Non-degenerate noise:*** *The diffusion matrix $\Sigma \in \mathbb{R}^{D\times D}$ is invertible.*
(iii) ***Finite energy:*** *For every finite $T > 0$, $\mathbb{E}\int_0^T \|G(X_t, X_{t-\tau_1}, \ldots, X_{t-\tau_q})\|^2\, dt < \infty$.*

To guarantee that the problem is well-posed, we need to control the drift locally and globally. Locally, it must be regular enough for trajectories to be unique, and globally it must not diverge in finite time to infinity. We thus impose the drift to be locally Lipschitz — a condition generally satisfied in most physical systems — and a radial Khasminskii condition (Khasminskii, 2012) which requires that, outside a ball whose radius can be as large as desired, the drift does not push the current state away from c faster than linearly and prevents explosiveness in finite time. It is much weaker than *dissipativity*, which asks the drift to actively pull the state back towards c when it is far away, like friction or damping in a physical system.

Assumption 2 (Regularity)**.** *To guarantee well-posedness and stability, we assume:*

(i) ***Local Lipschitzness:*** *$\{f_j\}_{j\leq D}$ are locally Lipschitz continuous in all of their arguments jointly.*
(ii) ***Radial Khasminskii condition:*** *There exist constants $c \in \mathbb{R}^D$, $R > 0$, and $\beta > 0$ such that, writing $z = (x_0, x_1, \ldots, x_q) \in \mathbb{R}^{D(q+1)}$ and $c^{(q)} := (c, \ldots, c) \in \mathbb{R}^{D(q+1)}$ for the concatenation of $q+1$ copies of c, the true drift G satisfies*

$$\langle x_0 - c, G(z)\rangle \leq \beta(1 + \|z - c^{(q)}\|^2) \quad \textit{for all } z \textit{ with } \|z - c^{(q)}\| \geq R. \tag{2}$$

These two sets of assumptions are permissive and encompass a wide range of SDDE-driven systems. We now establish the existence of a unique strong solution that does not explode in finite time, deferring the proof to Section A.2.

Proposition 1 (Non-explosion and strong solution)**.** *Under Assumptions 1 and 2, the SDDE defined by Equation 1 has a unique strong solution for every initial condition with* $\mathbb{E}\|X_0^{\mathrm{seg}}\|_{\mathcal{C}}^2 < \infty$ *without explosion i.e. it does not diverge to infinity in a finite amount of time.*

3.2 Identifiability of the dynamical drivers

We aim to identify the drift on the regions of the state space that the process actually visits. Since the drift depends on the lagged state Z_t and we observe a trajectory over $[0, T]$, the natural measure of "where the data lives" is the time-averaged law of Z_t. We thus define the *joint occupation measure* of the states entering the drift: for a Borel set $A \subseteq \mathbb{R}^{D(q+1)}$ and a fixed $T > \tau_{\max}$,

$$\bar{\mu}_T^q(A) \;:=\; \frac{1}{T - \tau_{\max}} \int_{\tau_{\max}}^{T} \mathbb{P}(Z_t \in A)\, dt. \tag{3}$$

A parameter of a model is identifiable if it is theoretically possible to uniquely determine it from observations. We thus consider an idealized learner: its model class must be able to represent the true system and minimize the population risk (Assumption 3). Although we never fully minimize the risk in practice, this is a standard assumption.

Assumption 3 (Model class and population risk minimization)**.** *Let* $T > \tau_{\max}$ *be fixed.*

(i) ***Model class:*** *The model class* $\mathcal{H}^q$ *consists of pairs* $(\widetilde{G}, \tilde{M})$ *of candidate masks* $\tilde{M} \in \{0,1\}^{D\times D\times(q+1)}$ *and masked drifts* $\widetilde{G}_j(x_0, x_1, \ldots, x_q) = \widetilde{f}_j(\tilde{m}_j^0 \odot x_0, \tilde{m}_j^1 \odot x_1, \ldots, \tilde{m}_j^q \odot x_q)$ *with continuous* $\widetilde{f}_j$*, and the* same known lags $\tau_1 < \cdots < \tau_q$ *as the data-generating process. It contains the true pair* (G, M)*, and every* $\widetilde{G} \in \mathcal{H}^q$ *has finite risk:*

$$\mathcal{R}_T(\widetilde{G}) \;:=\; \mathbb{E} \int_{\tau_{\max}}^{T} \left\|\widetilde{G}(Z_t) - G(Z_t)\right\|^2 dt = (T - \tau_{\max})\|\tilde{G} - G\|^2_{L^2(\bar{\mu}_T^q)} \;<\; \infty. \tag{4}$$

(ii) ***Risk minimization:*** *The learned pair* $(\widetilde{G}, \tilde{M}) \in \mathcal{H}^q$ *minimises* $\mathcal{R}_T$ *over* $\mathcal{H}^q$*.*

When minimizing the risk, an input along which G_j is constant could be kept or dropped without changing the drift, we therefore require every true edge to actually vary (Assumption 4).

Assumption 4 (Faithfulness)**.** *For every* $a \in \{0, 1, \ldots, q\}$ *and every pair* (i, j) *with* $M_{ij}^a = 1$*, there exist* $z, z' \in \mathrm{supp}(\bar{\mu}_T^q) = \mathbb{R}^{D(q+1)}$ differing only in the i-th coordinate of the a-th block argument *such that* $G_j(z) \neq G_j(z')$*. We show that* $\mathrm{supp}(\bar{\mu}_T^q) = \mathbb{R}^{D(q+1)}$ *in Lemma 1.*

Conversely, to avoid retaining spurious edges on which $\widetilde{G}_j$ does not actually depend, we bound the number of learned edges. This entails that the sparsity level of the system is known a priori, a standard assumption in identifiability (Lachapelle et al., 2023), and we relax it in Section 3.3.

Assumption 5 (Structural Sparsity)**.** *The learned mask is no denser than the true mask:*

$$\|\tilde{M}\|_0 \leq \|M\|_0. \tag{5}$$

Theorem 1 (Identifiability)**.** *Under Assumptions 1–5 , G and M are identifiable:*

$$\widetilde{G} = G \quad \text{on } \mathbb{R}^{D(q+1)}, \qquad \tilde{M} = M. \tag{6}$$

The proof of this theorem is given in Section A.3.

3.3 L_0-penalized identifiability

We now relax Assumption 5: instead of a hard constraint on the number of edges, it is possible to use a penalty if dropping a true edge to minimize sparsity leads to a higher increase in risk. Define, for each lag $a \in \{0, \ldots, q\}$ and pair (i, j) with $M_{ij}^a = 1$, the per-link *signal strength*

$$s_{ij}^a \;:=\; \left\| G_j - \Pi_{-(i,a)} G_j \right\|_{L_2(\bar{\mu}_T^q)}, \tag{7}$$

where $\Pi_{-(i,a)}$ is the $L_2(\bar{\mu}_T^q)$-orthogonal projection onto the closed subspace of functions of all coordinates *except* coordinate i of block a. In words, s_{ij}^a is the smallest error achievable by any model that ignores this input. Set

$$s_{\min} := \min\{s_{ij}^a : M_{ij}^a = 1\}, \qquad d_{\max} := \max_j \sum_{a=0}^{q} \|m_j^a\|_0. \tag{8}$$

Define, for a constant $\lambda > 0$, the penalized risk:

$$J_\lambda(G, M) = \mathcal{R}_T(G) + \lambda \sum_{a=0}^{q} \|M^a\|_0 \tag{9}$$

Theorem 2 (Identifiability with the penalized objective)**.** *Let Assumptions 1, 2 and 3(i) hold. Suppose $M \neq 0$ and $s_{\min} > 0$ i.e. the system has at least one true driver, and every true driver has a measurable impact on the dynamics. This replaces Assumption 4 (faithfulness). If*

$$0 < \lambda < \frac{(T - \tau_{\max}) s_{\min}^2}{d_{\max}}, \tag{10}$$

then every minimizer of J_λ over $\mathcal{H}^q$ satisfies $\tilde{M} = M$ and $\widetilde{G} = G$ on $\mathbb{R}^{D(q+1)}$. Moreover, every feasible pair with $\tilde{M} \neq M$ satisfies

$$J_\lambda(\tilde{G}, \tilde{M}) - J_\lambda(G, M) \geq \gamma(\lambda) := \min(\lambda, (T - \tau_{\max}) s_{\min}^2 - \lambda d_{\max}) > 0$$

Note that Assumption 5 (sparsity constraint) is not needed here.

3.4 Ergodicity of the true process

The results above assume that we can compute and minimize $\mathcal{R}_T$. In practice, however, we want to learn the system from discrete observations of a single long trajectory. To this aim, we need the system to be *ergodic*, so that trajectories converge to a unique statistical equilibrium. This is the natural notion of predictability for chaotic systems: individual trajectories cannot be forecast over long horizons, but their long-run statistics are well defined and do not depend on initial conditions.

Assumption 6 (Delay-adapted dissipativity)**.** *There exist $c \in \mathbb{R}^D$, $R, \alpha, C_R \geq 0$ and $\beta_0, \ldots, \beta_q, \varepsilon_1, \ldots, \varepsilon_q \geq 0$ such that, for all $z = (x_0, \ldots, x_q) \in \mathbb{R}^{D(q+1)}$,*

$$\langle x_0 - c, G(z) \rangle \leq \alpha - \beta_0 \|x_0 - c\|^2 + \textstyle\sum_{a=1}^q \beta_a \|x_a - c\|^2 \qquad \text{if } \|x_0 - c\| \geq R, \tag{11}$$

$$\langle x_0 - c, G(z) \rangle \leq C_R + \textstyle\sum_{a=1}^q \varepsilon_a \|x_a - c\|^2 \qquad \text{if } \|x_0 - c\| \leq R, \tag{12}$$

$$B := \textstyle\sum_{a=1}^q (\beta_a + \varepsilon_a) < \beta_0. \tag{13}$$

Far from c, Equation 11 forces the current state to be pulled back at rate β_0, while past states may push outwards at rates β_a. Near c, Equation 12 only forces the drift to stay radially controlled: no restoring condition is imposed on the dynamics inside the ball, which can be chaotic. Strict domination (Equation 13) is the delay counterpart of stability: the restoring force of the present must dominate the destabilising influence of the past. This assumption holds for most damped systems with a bounded attractor (Lorenz, 1963; 1996), for any bounded delayed feedback (Mackey & Glass, 1977), and for any linear delayed feedback once the instantaneous damping is superlinear (e.g. a delayed oscillator model of ENSO (Suarez & Schopf, 1988)).

Theorem 3 (Ergodicity: existence, uniqueness, exponential convergence)**.** *Under Assumptions 1, 2 and 6, the segment process admits a unique invariant probability measure π^{seg} on $\mathcal{C}$. Write $P_t(\varphi, \cdot) := \mathcal{L}(X_t^{\text{seg}} \mid X_0^{\text{seg}} = \varphi)$ and $V(\varphi) := \|\varphi - c\|_{\mathcal{C}}^2$, where c denotes the constant segment with value c. The invariant measure satisfies*

$$\int_{\mathcal{C}} \|\eta - c\|_{\mathcal{C}}^{2p} \, \pi^{\text{seg}}(d\eta) < \infty, \qquad p \geq 1 \text{ an integer}. \tag{14}$$

There exist $C \geq 1$ and $\gamma > 0$ such that, $\forall \epsilon > 0$, the metric $d(\varphi, \psi) := 1 \wedge \epsilon^{-1} \|\varphi - \psi\|_{\mathcal{C}}$ satisfies

$$\mathcal{W}_d\big(P_t(\varphi, \cdot), \pi^{\text{seg}}\big) \leq \|P_t(\varphi, \cdot) - \pi^{\text{seg}}\|_{\text{TV}} \leq C e^{-\gamma t}\big(1 + V(\varphi)\big), \qquad t \geq 0, \quad \varphi \in \mathcal{C}. \tag{15}$$

In words, the system effectively forgets about its initial condition, and reaches exponentially quickly a regime that is uniquely determined and well behaved.

3.5 FINITE-SAMPLE IDENTIFIABILITY

This section characterizes identifiability when only a finite number of samples are available. We assume data is observed on a deterministic time grid $0 = t_0 < t_1 < \cdots < t_n = T$, with step sizes $\Delta_k := t_{k+1} - t_k$ bounded by $\Delta_{\max} := \max_{k<n} \Delta_k \leq 1$. We also assume that the lagged connections happen within the observed time steps. Consequently, we restrict our analysis to the effective horizon $T_\circ := T - \tau_{\max} > 0$, denoting the sampled sequence as $Z_k := Z_{t_k}$. All subsequent summations over $k < n$ apply strictly to the usable window $t_k \geq \tau_{\max}$.

Assumption 7 (Finite-sample model class)**.** *Let $\mathcal{H}_n^q$ denote the restricted class of deterministic models derived from Assumption 3(i) satisfying the following two conditions:*

(i) ***Local finite covering:*** *For every radius $\rho \geq 1$ and tolerance $\epsilon > 0$, we can select a finite number of representative functions from within $\mathcal{H}_n^q$ (an internal ϵ-net) such that every function in the class is within a maximum distance of ϵ from at least one representative over the compact set $K_\rho := \{z \in \mathbb{R}^{D(q+1)} : \|z - c^{(q)}\| \leq \rho\}$. The minimal number of representatives required to cover the class this way is denoted by $N_n(\rho, \epsilon)$.*

(ii) ***Stationary envelope:*** *The pointwise diameter of the class, defined as $\mathcal{E}(z) := \sup_{\tilde{G}_1, \tilde{G}_2 \in \mathcal{H}_n^q} \|\tilde{G}_1(z) - \tilde{G}_2(z)\|$, has a finite second moment under the stationary measure: $m_{\mathcal{E}}^2 := \int \mathcal{E}(z)^2 \mu_\infty^q(dz) < \infty$.*

This ensures that each bounded region can be covered by a finite number of candidate functions (i), and that candidates do not lead to wildly different equilibrium (ii).

The empirical risk is defined below, and we assume that a measurable global minimizer exists:

$$\widehat{\mathcal{R}}_n(\tilde{G}) := \sum_k \Delta_k \left\| \frac{X_{t_{k+1}} - X_{t_k}}{\Delta_k} - \tilde{G}(Z_k) \right\|^2$$

$$(\widehat{G}, \widehat{M}) \in \operatorname*{arg\,min}_{(\tilde{G}, \tilde{M}) \in \mathcal{H}_n^q} \left\{ \widehat{\mathcal{R}}_n(\tilde{G}) + \lambda \sum_{a=0}^{q} \|\tilde{M}^a\|_0 \right\} \tag{16}$$

Theorem 4 (Finite-sample exact support recovery (L_0 estimator))**.** *Let Assumptions 1, 2, 6 and 7 hold, with $T > \tau_{\max}$. For a fixed $\delta \in (0, 1)$, there exists a bound $\varepsilon_n(\delta; \rho, \epsilon)$ dependent on two deterministic parameters ρ and ϵ, such that the uniform deviation is bounded with probability at least $1 - \delta$:*

$$\sup_{\tilde{G} \in \mathcal{H}_n^q} \left| \left[\widehat{\mathcal{R}}_n(\tilde{G}) - \widehat{\mathcal{R}}_n(G) \right] - \mathcal{R}_T(\tilde{G}) \right| \leq \varepsilon_n(\delta; \rho, \epsilon), \tag{17}$$

The explicit bound $\varepsilon_n(\delta; \rho, \epsilon)$ and details on the selection of ρ and ϵ are given in (56)–(58).

Suppose further that the true mask $M \neq 0$, the minimum signal strength $s_{\min} > 0$, and the penalty parameter λ satisfies $0 < \lambda < T_\circ s_{\min}^2 / d_{\max}$. This ensures a strictly positive margin $\gamma(\lambda) := \min\{\lambda, T_\circ s_{\min}^2 - \lambda d_{\max}\} > 0$.

If the error bound is strictly less than the margin ($\varepsilon_n(\delta; \rho, \epsilon) < \gamma(\lambda)$), every measurable minimizer of the empirical objective achieves exact support recovery with probability at least $1 - \delta$:

$$\mathbb{P}_\pi \left(\widehat{M}^a = M^a \text{ for all } a = 0, \ldots, q, \quad \|\widehat{G} - G\|_{L^2(\mu_\infty^q)}^2 \leq \frac{\varepsilon_n(\delta; \rho, \epsilon)}{T_\circ} \right) \geq 1 - \delta \tag{18}$$

Setting $\lambda = \lambda_0 T_\circ$ (where $0 < \lambda_0 < s_{\min}^2 / d_{\max}$) allows us to identify the mask exactly and the drift up to a given level of error from finitely many observations. Achieving consistency inherently requires joint conditions on the observation duration, mesh size, and model class complexity; increasing the number of samples alone is not sufficient. The bound ε_n is explicit, but it should be read as a qualitative consistency statement: it contains factors $1/\delta$ and a local envelope E_ρ that may grow with the radius ρ, itself growing with $T_\circ$.

3.6 IMPLEMENTATION AND OPTIMIZATION

The L_0-norm used in our identifiability theorems (1, 4) is not differentiable. To allow for differentiable optimization, we implement the relaxed L_0 penalization proposed by Louizos et al. (2018)

and parametrize the matrix M as the sigmoid of differentiable logit parameters. The functions $\widetilde{f}$ are parametrized with simple MLPs, with tanh activation functions and an affine output layer. With discrete, regularly sampled observations, we train the model to predict $X_{t+1} - X_t$ given $X_t, X_{t-1}, ..., X_{t-\tau}$. With irregular observations, it is possible to choose any numerical integrator, as when training Neural ODEs (Chen et al., 2018). Here, we use a simple Euler integrator, as all datasets in this paper have regular discrete temporal resolutions. We standardize each dimension by the estimated gradient mean and standard deviation. Otherwise, the method may learn to predict the mean for smooth dimensions, and mask off all of their potential drivers. We use the Adam optimizer (Kingma & Ba, 2015). We additionally propose to use gradient penalty (Gulrajani et al., 2017), to regularize our network and ensure that it does not overfit to spurious correlations. Full implementation details are provided in Section C.1. As the datasets considered in this paper are low-dimensional, and the architectures are relatively small, we run all experiments on CPUs. The code for implementing the methods can be found here `https://anonymous.4open.science/r/ODEDriverIdentifiability-C325/README.md`.

Assumption 6 requires the true drift to be unbounded, and Assumption 7 requires the drift to belong to the model class. However, MLPs with $\tanh$ activation functions are bounded by construction. To satisfy both assumptions needed for finite-sample identifiability, we parametrize $\widetilde{G}$ as $\widetilde{G}_R(z) = \widetilde{f}(z) - K \cdot (u - \Pi_{R,c}(u))$, where R is an arbitrarily large radius parameter, c the center of the system, $K = \mathrm{diag}(\kappa_1, \ldots, \kappa_D)$ a diagonal matrix, and $\Pi_{R,c}(u)$ a projection onto the cube of center c and radius R. In practice, we can choose R large enough so that the cube encompasses all our training observations, c to be equal to zero. This correction is seen as a linear restoring force outside a cube and can be applied at inference only. More details on this modification are given in A.9.

4 Datasets and Experiments

4.1 CausalDynamics Dataset

The CausalDynamics dataset (Herdeanu et al., 2025) contains more than 14,000 simulated dynamical systems grouped in different classes: "Simple" three-dimensional chaotic dynamical systems (e.g. the Lorenz system), with or without added noise (high noise level, 2) and confounders (585 graphs); "Coupled" systems generated from simulated interactions between simple MLPs, of dimension 3–10, with added noise and with or without confounders, lagged interactions, or standardization (14096 graphs); and 2 simulated "Climate" datasets, one corresponding to a simplified coupled ocean-atmosphere model in dimension 10, and one simulating the decoupled El-Niño Southern Oscillation (ENSO), an important mode of variability of the climate system (12 graphs).

For each graph, 10 timeseries of 1000 timesteps are generated with different random seeds. The area under the receiver-operator curve (AUROC), area under the precision-recall curve (AUPRC) and structural Hamming distance (SHD) are computed for each time series, before being averaged across the system class (e.g. "Simple, Confounder, No noise"). Hyperparameter search is done for all evaluated methods by maximizing AUROC on the "Simple, No confounder, No noise, Lorenz84" graph, which is then excluded from the results. The same set of parameters is used for evaluation on all other graphs. We implement L_0-NDDE and additional baselines: AGL (Bellot et al., 2022; Wu et al., 2024) and C-NODE (Aliee et al., 2025), which uses the L_1-penalty. We extend C-NODE to SDDEs similarly as Neural ODEs, by adding neural networks to capture lagged connections. All methods share the same core architecture. Hyperparameter search is conducted for each method separately, following the benchmark procedure. Details on the hyperparameter search, and the final set of parameters are reported in Section C.

4.2 The Dysts Benchmark

To evaluate the quality of the learned dynamics, we use the Dysts benchmark (Gilpin, 2021). It allows us to generate trajectories for low-dimensional chaotic systems (e.g. "Lorenz" system), which we use to evaluate L_0-NDDE, C-NODE, AGL, as well as PathReg, which uses L_0 to regularize entire input-output paths, and regular NODEs. In total, we use the 40 systems for which the Jacobian is given by the dataset, which is needed for our metric computation (Section E.2). For each system, we generate one noisy training trajectory of 50 Lyapunov timesteps, with stochastic noise forcing of

level 0.05, and 3 test trajectories of 250 Lyapunov timesteps, with different initial conditions. The Lyapunov timestep is the timescale over which a chaotic system loses its predictability (Lyapunov, 1992). These trajectories are long enough to densely cover the system's space. We train each method on the training trajectory, generate 3 trajectories using the 3 test initial conditions, and compare the predicted trajectories to the true unseen test trajectories for evaluation.

We compute two metrics to evaluate short-term accuracy: the normalized root mean-square error (NRMSE) averaged over the first 5 Lyapunov timesteps, and the valid Lyapunov prediction time (VLPT), the number of Lyapunov timesteps before the predicted trajectories diverge from the true trajectory by more than 1 standardized unit. We compute two metrics to evaluate the reconstructed systems' attractor geometry and chaoticity: the first Lyapunov exponent squared error (FLEE), indicative of the level of chaoticity of the system, and the squared error between the true and estimated Kaplan-Yorke dimension (D_{KY}, Evans et al., 2000), a proxy for the system's attractor dimension. We compute two metrics to evaluate the long-term statistics of the predicted trajectories: the log-spectral distance (LSD), to evaluate whether the oscillation frequencies are accurately captured, and the state-space distance (D_{stsp}, Koppe et al., 2019), a divergence between the true and estimated equilibrium distribution. We report, for each metric and method, the median value across all datasets and the average fractional rank, to assess whether one method systematically outperforms another. More details on the metric computation are given in Section E.2.

We choose 5 systems at random for hyperparameter search and we select, for each method, the set of parameters that maximizes the VLPT before evaluating the method on the other systems. A detailed description of the hyperparameter search and final parameters is given in Section E.1.

5 Results

5.1 CausalDynamics Benchmark

Results on the CausalDynamics benchmark are reported in Table 1. L_0-NDDE achieves lower average rank on all three metrics. It generally has higher AUROC/AUPRC than competing methods on the "Simple", "Coupled" and the "Climate" datasets, and consistently achieves low SHD, although the hyperparameters, only selected on the "Simple lorenz84" dataset, lead to too dense graphs on the "Coupled" datasets. Some results lead to "degenerate" metrics (e.g. AUROC of 0.50), as some methods sometimes fully sparsify or do not sparsify at all, and thus do not rank the edges. NGC and AGL achieve the lowest SHD on the MAOOAM dataset despite low AUROC, as they do not sparsify the learned graph. C-NODE performs well in general, despite not providing any identifiability guarantee. C-NODE also does not explicitly remove input features and even when a low weight is given to an input feature, the neural network can compensate for it.

5.2 Dysts Benchmark

Results are reported in Table 2. L_0-NDDE and NODE consistently rank better than other methods in all metrics except the NRMSE. The median D_{stsp} is extremely high for AGL and PathReg, as this metric explodes when the two distributions are very different. It is an artefact of D_{stsp}. Additionally, we show predicted versus true trajectories for L_0-NDDE, NODE, and C-NODE on 26 test systems, in Figure 2. Both NODE and L_0-NDDE are capable of modelling the systems very well.

To explain why NODE and L_0-NDDE outperform other methods, we report in Figure 3 the relationship between each metric and the SHD (upper half) and training MSE (lower half). We run a Kendall-τ test (Kendall, 1938), standard for non-Gaussian low-samples regimes, to measure the strength of association between the two sets of values. We find that D_{KY}, D_{stsp} and VLPT have a statistically significant association with both SHD and training MSE, indicating that on this benchmark, methods that better learn the drivers, and better fit the training data perform better. C-NODE, PathReg, and AGL all apply a regularization on the weights of the neural network, preventing these methods from reaching low training error. PathReg tends to sparsify specific input-output paths without masking off input features, thus not performing input feature selection. This is also why we do not evaluate it on CausalDynamics. With no weight penalty, neural ODE achieves low training error and thus good reconstruction performance. L_0-penalty does not penalize the weights of the network but instead turns the inputs on/off, allowing L_0-NDDE to achieve low train error while masking off irrelevant input features. With the lowest SHD, L_0-NDDE thus outperforms other methods.

	Experiment	Simple			Coupled					Climate		Avg. rank
		Default	Confounder	Noise	Default	Noise	Confounder	Time-lag	Standardize	MAOOAM	ENSO	
SHD	PCMCI+	41.04	23.02	47.64	224.80	183.90	324.63	327.72	228.32	80.00	529.36	7.30
	F-PCMCI	35.30	21.07	45.09	192.90	149.60	195.74	350.61	201.79	130.00	530.27	5.80
	NGC	28.91	19.96	28.84	840.95	842.55	670.53	793.67	840.26	31.00	337.09	7.90
	CUTS+	48.11	24.04	61.32	**152.00**	150.50	272.68	247.22	310.63	130.00	608.73	7.50
	RCD	61.85	26.74	61.38	157.05	155.65	136.53	**201.11**	**159.84**	130.00	665.36	6.80
	GRaSP	59.04	27.09	60.18	215.94	842.55	**136.00**	793.67	840.26	126.00	666.27	9.70
	AGL	26.60	16.96	21.28	319.04*	222.41*	246.37*	315.54*	312.31*	**30.00***	337.09	5.55
	C-NODE (L1)	**24.07**	17.10	**20.50**	253.91*	176.30*	179.58*	273.57*	258.09	42.00	335.36	4.50
	L_0**-NDDE (Ours)**	26.42	**16.25**	23.67	205.35	**144.20**	143.33	222.34	212.11	40.00	**325.54**	**2.50**
AUROC / AUPRC	PCMCI+	.52 / .71	.49 / .59	.50 / .69	**.67** / .25	.64 / .25	.58 / .20	.58 / .24	**.69** / .27	.69 / .88	.57 / .70	4.10 / 5.25
	F-PCMCI	.51 / .70	.50 / .59	.52 / .70	**.67** / .27	.57 / .21	.55 / .19	.59 / .24	.68 / .28	.50 / .81	.57 / .70	4.80 / 6.05
	NGC	.50 / .69	.50 / .58	.50 / .68	.50 / .15	.50 / .15	.50 / .16	.50 / .20	.50 / .15	.50 / .81	.50 / .67	9.85 / 10.90
	CUTS+	.50 / .69	.50 / .58	.50 / .68	.50 / .15	.50 / .15	.49 / .16	.50 / .20	.50 / .15	.50 / .81	.50 / .67	10.10 / 10.90
	RCD	.50 / .69	.50 / .58	.50 / .68	.50 / .15	.50 / .15	.51 / .18	.50 / .20	.50 / .16	.50 / .81	.50 / .67	9.55 / 10.25
	GRaSP	.52 / .71	.55 / .64	.49 / .68	.49 / .15	.50 / **.50**	.50 / .17	.50 / **.50**	.50 / **.50**	.48 / .81	.50 / .50	9.60 / 6.70
	AGL	.52 / .71*	.56 / .64	.53 / .74*	.53 / .33*	.53 / .33*	.54 / .35*	.52 / .35*	.54 / .34*	.56 / .86*	.54 / .69*	5.80 / 4.35
	C-NODE (L1)	.64 / .79	.54 / .65	.61 / .79*	.64 / **.43**	.63 / .43	**.67** / .47	.59 / .43	.68 / .47	**.88 / .97**	.49 / .66*	3.45 / 3.35
	L_0**-NDDE (Ours)**	**.64 / .80**	**.58 / .66**	**.63 / .80**	.64 / **.43**	**.65** / .44	**.67** / .47	**.60** / .43	.66 / .45	.86 / .96	**.58 / .75**	**1.75 / 2.10**

Table 1: L_0**-NDDE outperforms competing methods on the CausalDynamics benchmark.** SHD (lower is better) and AUROC / AUPRC (higher is better) are reported for L_0-NDDE (Ours), AGL, C-NODE and the benchmark causal discovery methods. The best values are bold, the second best values are underlined. * indicates statistical significance between AGL or C-NODE and L_0-NDDE, calculated with a Welch's t-test (p-value 0.05). The benchmark only provides the mean values for the causal discovery methods, and we thus do not run statistical significance tests for these methods. For AUROC and AUPRC, statistical significance results always match, so we report only one * for each pair. The last column reports the average rank for each method and metric. For readability, methods that never score best are not shown here, but in the Appendix (Table 4)

Method	NRMSE (↓)		VLPT (↑)		FLEE (↓)		D_{KY} (↓)		LSD (↓)		D_{stsp} (↓)		SHD (↓)	
	Med.	Rank	Med.	Rank	Med.	Rank	Med.	Rank	Med.	Rank	Med.	Rank	Med.	Rank
AGL	**1.15**	2.69	2.85	3.14	0.25	3.47	0.88*	3.82	14.71*	2.82	3e^6*	4.37	0.44*	4.46
PathReg	1.19	**2.49**	4.33	3.24	0.25	3.78	0.89*	3.9	16.99	3.02	3e^6*	4.62	0.22	2.70
C-NODE	1.53	4.05	4.61	3.39	0.20	3.00	0.72*	3.17	37.79*	4.08	34.60*	2.86	0.17	2.67
NODE	1.20	2.94	5.12	2.63	**0.16**	2.51	0.076	2.17	**13.99**	2.57	4.29	1.62	0.22	2.70
L_0-NDDE	1.19	2.83	**6.54**	**2.60**	**0.16**	**2.25**	**0.075**	**1.94**	14.75	**2.50**	**2.50**	**1.54**	**0.11**	**2.45**

Table 2: L_0**-NDDE outperforms other feature selection methods.** L_0-NDDE ranks best in five of the metrics. The difference with NODE is non significant, as NODE also performs very well, but L_0-NDDE consistently ranks better than the other methods performing input feature selection, especially on attractor dimension reconstruction (D_{KY}) and convergence in distribution (D_{stsp}). * indicates statistical significance with the L_0-NDDE metric, obtained using a permutation test (p-value 0.05). For each metric, best results are bolded and second best are underlined.

6 Discussion

We hope that the theoretical framework proposed in this paper lays the foundations for the development of robust methods for modelling lagged dynamical systems. One limitation of our method is that we can only optimize for an approximate L_0-penalty, as the true L_0-loss is not differentiable, although existing methods also rely on approximations. We show empirical improvements on two benchmarks that evaluate the methods mostly on SDE-driven systems as, to the best of our knowledge, no comprehensive benchmark exists for evaluating lagged systems modelling.

Despite these promising results, future work is needed to apply this method to real physical systems. Our method's computational requirements scale quadratically with the input dimension. To apply it to high-dimensional systems, it would need to be combined with an identifiable dimensionality reduction technique (Hı zlıet al., 2025). Our method learns probabilities for the input-output mapping, but the learned drift is deterministic. Extending our method to be fully probabilistic could allow us to better represent uncertainty during long-term autoregressive rollouts. We see several exciting applications of our method. Identifying correct drivers in dynamical systems could open the door to scientific discovery across several domains. Our method could also be integrated into physical models to robustly parametrize SDDE-driven subprocesses. It could also be used to stabilize gradient estimates in differentiable models of chaotic systems (Metz et al., 2022).

AI USE STATEMENT

In this work, we used generative AI tools for providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and implementing methods. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. Additionally, we used generative AI tools to identify relevant literature. We have reviewed all AI-assisted work. For the mathematical proofs, we first sketched the proofs ourselves before asking an LLM to review them and suggest corrections or improvements. We have reviewed every suggestion manually before incorporating them into the final version of the proofs. For implementing methods, we carefully verified and tested all code used for the experiments in this paper. All the identified literature and references were manually checked, and we read every paper that our work builds upon or takes inspiration from; e.g., we did not rely on an LLM-generated summary of other research papers to design this paper, or to decide whether to include or not include references. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The code to implement the methods has been shared in Section 3.6. This code uses poetry to manage the environment. Detailed parameters, as well as hyperparameter search procedures are given in the main text as well as the appendix. As mentioned in the main text, all experiments have been run on CPU, allowing for easy reproducibility i.e. without intensive computational requirements. The metric computation is described in details in the appendix (Section E.2). The data used in this paper is taken from peer-reviewed, publicly available benchmarks.

REFERENCES

Hananeh Aliee, Till Richter, Mikhail Solonin, Ignacio Ibarra, Fabian Theis, and Niki Kilbertus. Sparsity in continuous-depth neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 901–914, 2022. doi: 10.52202/068431-0066.

Hananeh Aliee, Fabian J. Theis, and Niki Kilbertus. Beyond predictions in neural odes: Identification and interventions, 2025. URL https://arxiv.org/abs/2106.12430.

Alexis Bellot, Kim Branson, and Mihaela van der Schaar. Neural graphical modelling in continuous-time: consistency guarantees and algorithms. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=SsHBkfeRF9L.

Richard C. Bradley. Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probability Surveys*, 2:107 – 144, 2005. doi: 10.1214/154957805100000104. URL https://doi.org/10.1214/154957805100000104.

Philippe Brouillard, Sebastien Lachapelle, Julia Kaltenborn, Yaniv Gurwicz, Dhanya Sridhar, Alexandre Drouin, Peer Nowack, Jakob Runge, and David Rolnick. Learning a spatial partitioning and its causal relations from temporal data. In Bijan Mazaheri and Niels Richard Hanson (eds.), *Proceedings of the Fifth Conference on Causal Learning and Reasoning*, volume 323 of *Proceedings of Machine Learning Research*, pp. 547–592. PMLR, 06–08 Apr 2026. URL https://proceedings.mlr.press/v323/brouillard26a.html.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016. doi: 10.1073/pnas.1517384113. URL https://www.pnas.org/doi/abs/10.1073/pnas.1517384113.

Kurt Butler, Daniel Waxman, and Petar M Djurić. Tangent space causal inference: Leveraging vector fields for causal discovery in dynamical systems. *Advances in Neural Information Processing Systems*, 37:120078–120102, 2024.

Luca Castri, Sariah Mghames, Marc Hanheide, and Nicola Bellotto. Enhancing causal discovery from robot sensor data in dynamic scenarios. In *Conference on Causal Learning and Reasoning*, pp. 243–258. PMLR, 2023.

Mickaël D. Chekroun and Honghu Liu. Effective reduced models from delay differential equations: Bifurcations, tipping solution paths, and enso variability. *Physica D: Nonlinear Phenomena*, 460:134058, 2024. ISSN 0167-2789. doi: https://doi.org/10.1016/j.physd.2024.134058. URL https://www.sciencedirect.com/science/article/pii/S0167278924000095.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Yuxiao Cheng, Lianglong Li, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. Cuts+: High-dimensional causal discovery from irregular time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11525–11533, 2024.

Yu. A. Davydov. Mixing conditions for markov chains. *Theory of Probability & Its Applications*, 18 (2):312–328, 1974. doi: 10.1137/1118033. URL https://doi.org/10.1137/1118033.

J. R. Dorfman. *An Introduction to Chaos in Nonequilibrium Statistical Mechanics*. 1999.

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. *Advances in neural information processing systems*, 32, 2019.

Mari Dahl Eggen and Alise Danielle Midtfjord. Delay-sde-net: A deep learning approach for time series modelling with memory and uncertainty estimates, 2023. URL https://arxiv.org/abs/2303.08587.

Denis J. Evans, E. G. D. Cohen, Debra J. Searles, and F. Bonetto. Note on the kaplan–yorke dimension and linear transport coefficients. *Journal of Statistical Physics*, 101, 2000. doi: 10.1023/A:1026449702528.

Michael Ghil and Valerio Lucarini. The physics of climate variability and climate change. *Rev. Mod. Phys.*, 92:035002, Jul 2020. doi: 10.1103/RevModPhys.92.035002. URL https://link.aps.org/doi/10.1103/RevModPhys.92.035002.

Michael Ghil, Mickaël D. Chekroun, and Eric Simonnet. Climate dynamics and fluid mechanics: Natural variability and related uncertainties. *Physica D: Nonlinear Phenomena*, 237(14):2111–2126, 2008. ISSN 0167-2789. doi: https://doi.org/10.1016/j.physd.2008.03.036. URL https://www.sciencedirect.com/science/article/pii/S0167278908001139. Euler Equations: 250 Years On.

William Gilpin. Chaos as an interpretable benchmark for forecasting and data-driven modelling. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=enYjtbjYJrf.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912791.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 5769–5779, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

M. Hairer, J. C. Mattingly, and M. Scheutzow. Asymptotic coupling and a general form of harris' theorem with applications to stochastic delay equations. *Probability Theory and Related Fields*, 149(1):223–259, 2011. doi: 10.1007/s00440-009-0250-6. URL https://doi.org/10.1007/s00440-009-0250-6.

Aristide Halanay. *Differential Equations: Stability, Oscillations, Time Lags*, volume 23 of *Mathematics in Science and Engineering*. Academic Press, 1966.

Jack K. Hale and Sjoerd M. Verduyn Lunel. *Introduction to Functional Differential Equations*. Applied Mathematical Sciences. Springer, New York, NY, 1 edition, 1993. ISBN 978-0-387-94076-2. doi: 10.1007/978-1-4612-4342-7.

Benjamin Herdeanu, Juan Nathaniel, Carla Roesch, Jatan Buch, Gregor Ramien, Johannes Haux, and Pierre Gentine. Causaldynamics: A large-scale benchmark for structural discovery of dynamical causal models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-1349. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/39a3f7dc00e4723a5ca7808b04fc4d4f-Paper-Datasets_and_Benchmarks_Track.pdf.

Çağlar Hı zlı, Çağatay Yı ldız, Matthias Bethge, ST John, and Pekka Marttinen. Identifying latent state transitions in non-linear dynamical systems. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 71286–71315, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b192f601432dff4e6777d2d06d52cfa5-Paper-Conference.pdf.

Sebastian Hickman, Ilija Trajković, Julia Kaltenborn, Francis Pelletier, Alexander T Archibald, Yaniv Gurwicz, Peer Nowack, David Rolnick, and Julien Boussard. Causal climate emulation with bayesian filtering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=YxPI1c5e09.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11 (56):1709–1731, 2010. URL http://jmlr.org/papers/v11/hyvarinen10a.html.

Kiyosi Itô. Stochastic integral. *Proceedings of the Imperial Academy*, 20(8):519 – 524, 1944. doi: 10.3792/pia/1195572786. URL https://doi.org/10.3792/pia/1195572786.

Kiyosi Itô. *On Stochastic Differential Equations*. Number 4 in Memoirs of the American Mathematical Society. American Mathematical Society, 1951. doi: 10.1090/memo/0004. URL https://doi.org/10.1090/memo/0004.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics. Springer New York, NY, 2 edition, 1991. ISBN 978-0-387-97655-6. doi: 10.1007/978-1-4612-0949-2.

Andrew Keane, Bernd Krauskopf, and Claire M. Postlethwaite. Climate models with delay differential equations. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(11):114309, 10 2017. ISSN 1054-1500. doi: 10.1063/1.5006923. URL https://doi.org/10.1063/1.5006923.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 06 1938. ISSN 0006-3444. doi: 10.1093/biomet/30.1-2.81. URL https://doi.org/10.1093/biomet/30.1-2.81.

Rafail Khasminskii. *Stochastic Stability of Differential Equations*. Stochastic Modelling and Applied Probability. Springer Berlin, Heidelberg, 2 edition, 2012. ISBN 978-3-642-23279-4. doi: 10.1007/978-3-642-23280-0.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Georgia Koppe, Hazem Toutounji, Peter Kirsch, Stefanie Lis, and Daniel Durstewitz. Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fmri. *PLOS Computational Biology*, 15(8):1–35, 08 2019. doi: 10.1371/journal.pcbi.1007263. URL https://doi.org/10.1371/journal.pcbi.1007263.

Sebastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18171–18206. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/lachapelle23a.html.

Wai-Yin Lam, Bryan Andrews, and Joseph Ramsey. Greedy relaxations of the sparsest permutation algorithm. In James Cussens and Kun Zhang (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1052–1062. PMLR, 01–05 Aug 2022. URL https://proceedings.mlr.press/v180/lam22a.html.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13557–13603. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/lippe22a.html.

Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2): 130–141, 1963.

Edward N. Lorenz. Predictability: A problem partly solved. In *Proceedings of the Seminar on Predictability*, volume 1, pp. 1–18, Reading, UK, 1996. ECMWF.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through l_0 regularization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1Y8hhg0b.

A. M. Lyapunov. The general problem of the stability of motion. *International Journal of Control*, 55(3):531–534, 1992. doi: 10.1080/00207179208934253. URL https://doi.org/10.1080/00207179208934253.

Michael C. Mackey and Leon Glass. Oscillation and chaos in physiological control systems. *Science*, 197(4300):287–289, 1977. ISSN 00368075, 10959203. URL http://www.jstor.org/stable/1744526.

Takashi Nicholas Maeda and Shohei Shimizu. Rcd: Repetitive causal discovery of linear non-gaussian acyclic models with latent confounders. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 735–745. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/maeda20a.html.

Xuerong Mao. *Stochastic Differential Equations and Applications*. Woodhead Publishing, 2 edition, 2007. ISBN 978-1-904275-34-3. doi: 10.1533/9780857099402. URL https://www.sciencedirect.com/science/book/9781904275343.

Hiroki Masuda. Ergodicity and exponential β-mixing bounds for multidimensional diffusions with jumps. *Stochastic Processes and their Applications*, 117(1):35–56, 2007. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2006.04.010. URL https://www.sciencedirect.com/science/article/pii/S0304414906000524.

Luke Metz, C. Daniel Freeman, Samuel S. Schoenholz, and Tal Kachman. Gradients are not all you need, 2022. URL https://arxiv.org/abs/2111.05803.

S. E. A. Mohammed. The infinitesimal generator of a stochastic functional differential equation. In W.N. Everitt and B.D. Sleeman (eds.), *Ordinary and Partial Differential Equations*, pp. 529–538, Berlin, Heidelberg, 1982. Springer Berlin Heidelberg. ISBN 978-3-540-39561-4.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019. ISSN 2504-4990. doi: 10.3390/make1010019. URL https://www.mdpi.com/2504-4990/1/1/19.

David Nualart. *The Malliavin Calculus and Related Topics*. Probability and Its Applications. Springer Berlin, Heidelberg, 2 edition, 2006. ISBN 978-3-540-28328-7. doi: 10.1007/3-540-28329-3.

Yongkyung Oh, Seungsu Kam, Dongyoung Lim, and Sungil Kim. Modeling irregular astronomical time series with neural stochastic delay differential equations. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, pp. 5068–5073, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720406. doi: 10.1145/3746252.3760805. URL https://doi.org/10.1145/3746252.3760805.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International conference on artificial intelligence and statistics*, pp. 1595–1605. Pmlr, 2020.

Jeongjin Park, Nicole Tianjiao Yang, and Nisha Chandramoorthy. When are dynamical systems learned from time series data statistically accurate? In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 43975–44008. Curran Associates, Inc., 2024. doi: 10.52202/079017-1396. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4dc57702c987e1e72f0dd2921edb5ded-Paper-Conference.pdf.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations, 2017. URL https://arxiv.org/abs/1711.10561.

J. O. Ramsay, G. Hooker, D. Campbell, and J. Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796, 2007. doi: https://doi.org/10.1111/j.1467-9868.2007.00610.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2007.00610.x.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets, 2020.

M. Scheutzow. Qualitative behaviour of stochastic delay equations with a bounded memory. *Stochastics*, 12(1):41–80, 1984. doi: 10.1080/17442508408833294. URL https://doi.org/10.1080/17442508408833294.

Ch. Skokos. *The Lyapunov Characteristic Exponents and Their Computation*, pp. 63–135. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-04458-8. doi: 10.1007/978-3-642-04458-8_2. URL https://doi.org/10.1007/978-3-642-04458-8_2.

Daniel W. Stroock and S. R. Srinivasa Varadhan. On the support of diffusion processes with applications to the strong maximum principle. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 3, pp. 333–359. University of California Press, 1972. URL https://api.semanticscholar.org/CorpusID:35508438.

Max J. Suarez and Paul S. Schopf. A delayed action oscillator for enso. *Journal of Atmospheric Sciences*, 45(21):3283 – 3287, 1988. doi: 10.1175/1520-0469(1988)045⟨3283:ADAOFE⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/atsc/45/21/1520-0469_1988_045_3283_adaofe_2_0_co_2.xml.

George Sugihara, Robert May, Hao Ye, Chih hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *Science*, 338(6106):496–500, 2012. doi: 10.1126/science.1227079. URL https://www.science.org/doi/abs/10.1126/science.1227079.

Nicholas Tagliapietra, Katharina Ensinger, Christoph Zimmer, and Osman Mian. Causal structure learning for dynamical systems with theoretical score analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36740–36748, 2026.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B. Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2022. doi: 10.1109/TPAMI.2021.3065601.

Silviu-Marian Udrescu, Andrew Tan, Jiahai Feng, Orisvaldo Neto, Tailin Wu, and Max Tegmark. Ai feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4860–4871. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/33a854e247155d590883b93bca53848a-Paper.pdf.

Peter Walters. *An Introduction to Ergodic Theory*. Springer New York,, 1999. URL https://link.springer.com/book/9780387951522.

Philippe Wenk, Gabriele Abbati, Michael A Osborne, Bernhard Schölkopf, Andreas Krause, and Stefan Bauer. Odin: Ode-informed regression for parameter and state inference in time-continuous dynamical systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6364–6371, 2020.

Fan Wu, Woojin Cho, David Korotky, Sanghyun Hong, Donsub Rim, Noseong Park, and Kookjin Lee. Identifying contemporaneous and lagged dependence structures by promoting sparsity in continuous-time neural networks. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pp. 2534–2543, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3679751. URL https://doi.org/10.1145/3627673.3679751.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RDlLMjLJXdq.

Yuan Yin, Vincent Le Guen, Jérémie Dona, Emmanuel De Bézenac, Ibrahim Ayed, Nicolas Thome, and Patrick Gallinari. Augmenting physical models with deep networks for complex dynamics forecasting. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124012, 2021. doi: 10.1088/1742-5468/ac3ae5. URL https://doi.org/10.1088/1742-5468/ac3ae5.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994. ISSN 00911798, 2168894X. URL http://www.jstor.org/stable/2244496.

Han Zhang, Xi Gao, Jacob Unterman, and Tom Arodz. Approximation capabilities of neural ODEs and invertible residual networks, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/zhang20h.html.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric Xing. Dags with no tears: Continuous optimization for structure learning, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. doi: 10.1198/016214506000000735. URL https://doi.org/10.1198/016214506000000735.

Coşkun Çetin, Jose Roberto Castilho Piqueira, Burhaneddin İzgi, Ayse Peker-Dobie, Semra Ahmetolan, and Murat Özkaya. Deterministic, stochastic, and mean-field pde models in neuroscience. *Frontiers in Computational Neuroscience*, Volume 20 - 2026, 2026. ISSN 1662-5188. doi: 10.3389/fncom.2026.1762692. URL https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2026.1762692.

A PROOFS

We dedicate this section to proving the different lemmas, propositions and theorems from the main content of the paper as well as some auxiliary lemmas and corollaries. We first establish that the model is well posed: Proposition 1 gives a unique, non-exploding strong solution. We then treat identifiability at the population level, first for the exact-risk formulation (Theorem 1) and then for the penalized objective (Theorem 2). The remaining three subsections provide the probabilistic machinery to go from the population result to the finite-sample one: uniform moment bounds, ergodicity of the true process, and exponential mixing of the sampled lagged state. These feed the finite-sample recovery guarantee (Theorem 4). We provide a graph illustrating the relationships between the assumptions, lemmas, propositions, and theorems in Figure 1.

A.1 NOTATION

We denote the true data-generating objects using standard symbols: G for the drift, G_j for its coordinates, f_j for the link functions, and M for the binary mask. Candidates within the model class are indicated by a tilde (e.g., $\tilde{G}$), whereas empirical minimizers are denoted by a hat (e.g., $\hat{G}$). Matrix indices follow the mask variables, where $M^a_{ij} = 1$ indicates that the driver coordinate i at lag τ_a influences the target coordinate j. The lag index $a \in \{0, \dots, q\}$ specifies the lag block, with $a = 0$ representing the instantaneous interactions. We let m^a_j denote the j-th column of M^a, and define $S_j = \{(i, a) : M^a_{ij} = 1\}$ as the active set of drivers for coordinate j. The delay sequence is strictly ordered as $0 < \tau_1 < \cdots < \tau_q$, with the maximum delay defined as $\tau_{\max} = \tau_q$.

Because the drift evaluates the trajectory at multiple past times, the process X is non-Markovian, though the process of its recent histories (segments) is Markovian. We therefore operate on the *segment space* of continuous functions from $[-\tau_{\max}, 0]$ to $\mathbb{R}^D$,

$$\mathcal{C} = C\big([-\tau_{\max}, 0]; \mathbb{R}^D\big), \qquad X^{\text{seg}}_t(s) = X_{t+s}, \quad s \in [-\tau_{\max}, 0],$$

equipped with the supremum norm $\|\cdot\|_{\mathcal{C}}$. Generic segments in this space are denoted by Greek letters such as φ, ψ, and η, with X^{seg}_0 serving as the initial condition. The drift relies strictly on $q + 1$ discrete time slices of a segment; thus, we define the *lagged state* as $Z_t = (X_t, X_{t-\tau_1}, \dots, X_{t-\tau_q})$. Observations of this state are taken on a temporal grid $t_0 < \cdots < t_n = T$, where we write $Z_k = Z_{t_k}$. The effective horizon over which the drift is evaluated is $T_\circ = T - \tau_{\max}$. Reference balls centered at the constant c (introduced in our regularity and dissipativity assumptions) are denoted by $B_R \subset \mathcal{C}$ in the segment space and $K_\rho \subset \mathbb{R}^{D(q+1)}$ in the lagged state space, and $c^{(q)} \in \mathbb{R}^{D(q+1)}$ denotes the concatenation of $q + 1$ copies of c. When $q = 0$ (no delays) we use the conventions $\tau_{\max} = 0$ and $\mathcal{C} = \mathbb{R}^D$, so that the segment process is X itself.

We use subscripts on the probability operators to distinguish the initial state of the process. The probability measure and expectation for a solution initialized at a fixed segment $X^{\text{seg}}_0 = \varphi$ are written as $\mathbb{P}_\varphi$ and $\mathbb{E}_\varphi$, with the corresponding transition kernel $P_t(\varphi, \cdot)$ and semigroup $(P_t)_{t\geq 0}$. This specific notation is utilized for moment and coupling estimates that require uniform bounds over $\varphi \in B_R$. Conversely, when the process is initialized from equilibrium ($X^{\text{seg}}_0 \sim \pi^{\text{seg}}$, independent of the driving noise W), we may use $\mathbb{P}_\pi$ and $\mathbb{E}_\pi$. Here, π^{seg} represents the invariant measure of the system, which forms the probabilistic foundation for our finite-sample analysis. Undecorated expectations $\mathbb{E}$ and probabilities $\mathbb{P}$ denote an arbitrary initial distribution consistent with the underlying data-generating process.

We compare candidate drifts in the L^2 space associated with the joint occupation measure $\bar{\mu}^q_T$, which represents the time-averaged law of Z_t over the interval $[\tau_{\max}, T]$. Identifiability is first established as an identity in $L^2(\bar{\mu}^q_T)$ before being extended to all of $\mathbb{R}^{D(q+1)}$. Under stationary conditions, $\bar{\mu}^q_T$ coincides with the stationary law μ^q_∞ of Z_t, rendering our bounding constants independent of the horizon T.

For a given candidate drift $\tilde{G}$ and mask $\tilde{M}$, the population risk is defined as $\mathcal{R}_T(\widetilde{G}) = T_\circ \|\widetilde{G} - G\|^2_{L^2(\bar{\mu}^q_T)}$. Its empirical counterpart is the Euler criterion $\widehat{\mathcal{R}}_n$, leading to the penalized objective $J_\lambda(\widetilde{G}, \tilde{M}) = \mathcal{R}_T(\widetilde{G}) + \lambda \sum_a \|\tilde{M}^a\|_0$. Because $\widehat{\mathcal{R}}_n$ contains a quadratic-variation term identical across all candidates, we evaluate only the excess empirical risk $\widehat{\mathcal{R}}_n(\widetilde{G}) - \widehat{\mathcal{R}}_n(G)$ against

the population risk $\mathcal{R}_T$. Although $\mathcal{R}_T$ involves the unknown G, it is determined by observable quantities up to a candidate-independent constant: for every $\widetilde{G}$ with finite risk, the contrast $\mathcal{C}_T(\widetilde{G}) := \mathbb{E}\big[\int_{\tau_{\max}}^{T} \|\widetilde{G}(Z_t)\|^2\, dt - 2\int_{\tau_{\max}}^{T} \langle \widetilde{G}(Z_t), dX_t\rangle\big]$ satisfies $\mathcal{C}_T(\widetilde{G}) - \mathcal{C}_T(G) = \mathcal{R}_T(\widetilde{G})$, as follows by substituting the SDDE and using that the square-integrable stochastic integral has zero mean.

Finally, successful recovery is governed by three primary structural quantities. The first is the signal strength $s_{ij}^a = \|G_j - \Pi_{-(i,a)} G_j\|_{L^2(\bar{\mu}_T^q)}$, where $\Pi_{-(i,a)}$ denotes the orthogonal projection onto the subspace of functions that do not depend on (i.e. do not use) coordinate i in lag block a. The remaining two are $s_{\min}$, the minimum signal strength across all true edges, and $d_{\max}$, the maximal in-degree of the system. These combine to define the margin $\gamma(\lambda) = \min\{\lambda, T_\circ s_{\min}^2 - \lambda d_{\max}\}$, which acts as the minimum objective gap incurred by any incorrectly identified mask.

A.2 Non-explosion and existence of a strong solution

Since the process is not Markov, we cannot apply the classical existence theorems for SDEs directly. Instead, we exploit the fact that the delays are bounded away from zero and prove existence of a unique solution by induction over intervals of length $\tau_{\min} := \min_a \tau_a$ (the so-called method of steps, see e.g. Hale & Verduyn Lunel, 1993). On each such interval the delayed arguments are already known from the previous interval, so that the SDDE reduces to a standard (non-delayed) Itô equation with a random, time-dependent drift: we solve that by truncation on a closed ball and rule out explosion with a Khasminskii-type Lyapunov test. The induction then extends the solution to all of $[0, \infty)$. The proof of the following statement only uses the initial-condition clause of Assumption 1 (independence of X_0^{seg} and W, and its finite second moment) together with Assumption 2.

Proposition 1 (Non-explosion & Strong solution)**.** *Under Assumptions 1 and 2, the SDDE defined in Equation 1 has a unique strong solution for every initial condition with $\mathbb{E}\|X_0^{\mathrm{seg}}\|_{\mathcal{C}}^2 < \infty$ without explosion i.e. it does not diverge to infinity in a finite amount of time.*

Proof. Step 1: Initialisation on $[-\tau_{\max}, \tau_{\min}]$. The initial segment X_0^{seg} provides the solution on $[-\tau_{\max}, 0]$. For $s \in [0, \tau_{\min}]$ and every lag $a \geq 1$, the delayed argument satisfies $s - \tau_a \leq \tau_{\min} - \tau_a \leq 0$, so $X_{s-\tau_a}$ is read from the initial segment. Define

$$Y_s = (X_{s-\tau_1}, \ldots, X_{s-\tau_q}), \quad s \in [0, \tau_{\min}]$$

This is a continuous, $\mathcal{F}_0$-measurable path with compact range $K_Y \subset \mathbb{R}^{Dq}$ (as a continuous function on the compact $[0, \tau_{\min}]$ composed with the continuous initial segment). On this segment, the SDDE reduces to a standard (non-delayed) Itô equation:

$$dX_s = b(s, X_s)\, ds + \Sigma\, dW_s, \quad X_0 = X_0^{\mathrm{seg}}(0)$$

with time-dependent drift $b(s, x) := G(x, Y_s)$.

(a) Existence and uniqueness via truncated approximation. For $n \in \mathbb{N}$, consider the metric projection onto the closed ball of radius n centered at c, $\pi_n : \mathbb{R}^D \to \bar{B}_n(c)$ which is 1-Lipschitz by convexity. Define the *truncated drift*:

$$b_n(s, x) := b(s, \pi_n(x)) = G(\pi_n(x), Y_s)$$

Since G is locally Lipschitz by Assumption 2 and both π_n, Y_s remain in a compact set ($\bar{B}_n(c) \times K_Y$), the truncated drift b_n is globally Lipschitz in x uniformly in $s \in [0, \tau_{\min}]$ with Lipschitz constant L_n (random but $\mathcal{F}_0$-measurable).

Since L_n and Y are $\mathcal{F}_0$-measurable while the Brownian motion W is independent of $\mathcal{F}_0$ (Assumption 1), we may apply the Picard-Lindelöf theorem for SDEs with globally Lipschitz coefficients (Mao, 2007, Ch. 2) conditionally on $\mathcal{F}_0$, i.e. for each fixed realisation of the initial segment (alternatively, one may localise on the events $\{L_n \leq m\} \in \mathcal{F}_0$, $m \in \mathbb{N}$). Hence the truncated equation

$$dX_s^{(n)} = b_n(s, X_s^{(n)})\, ds + \Sigma\, dW_s, \quad X_0^{(n)} = X_0^{\mathrm{seg}}(0)$$

admits a unique non-exploding strong solution on $[0, \tau_{\min}]$.

Define the exit time from $\bar{B}_n(c)$, $\zeta_n := \inf\{s \geq 0, \|X_s^{(n)} - c\| \geq n\}$. For any $n \leq m$, the pathwise uniqueness on $[0, \zeta_n]$ gives $X_s^{(n)} = X_s^{(m)}$ for all $s \leq \zeta_n$ (since $b_n = b_m = b$ on the ball $\bar{B}_n(c)$).

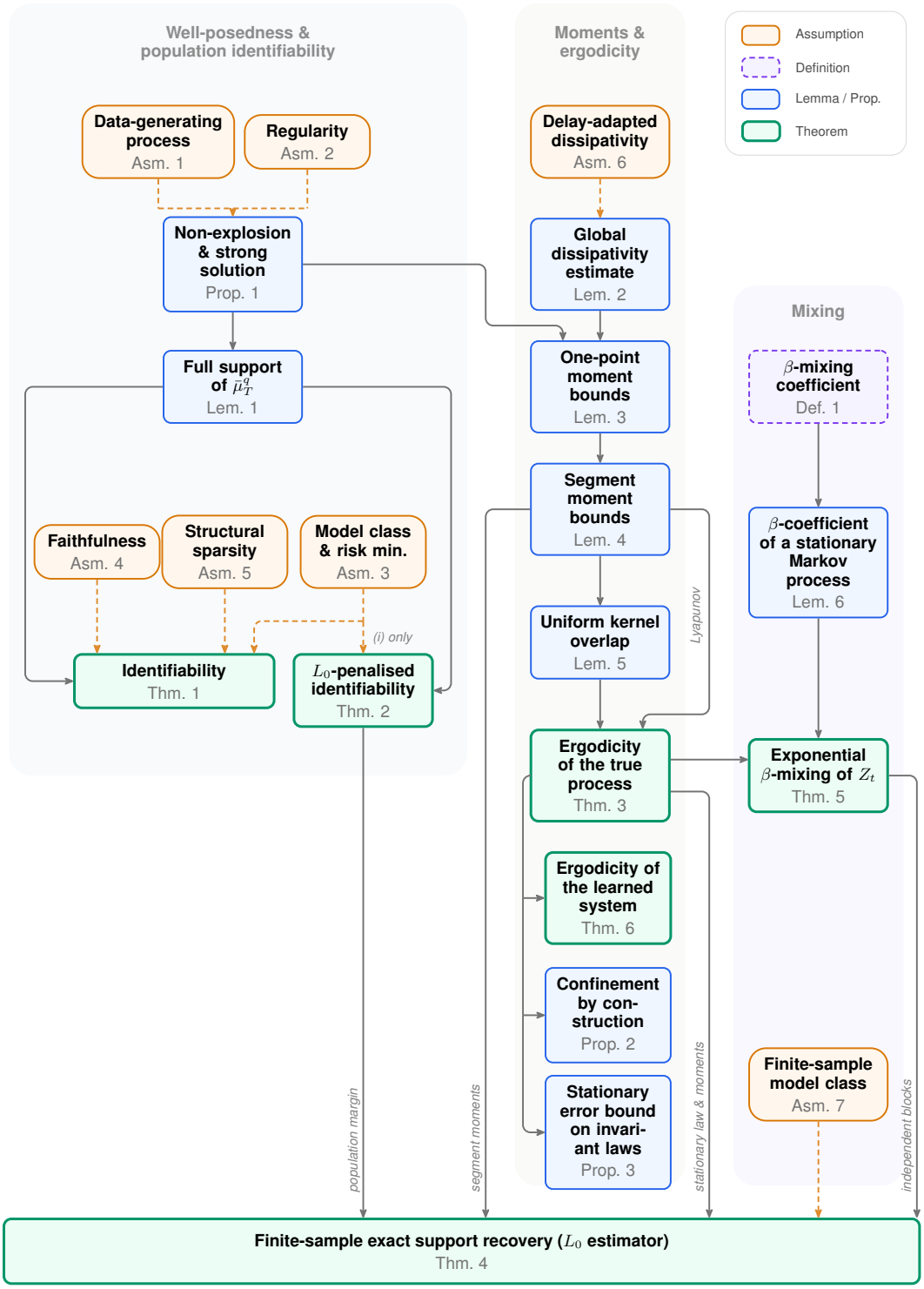

Figure 1: Dependency map of the theoretical results. Yellow boxes are assumptions, blue boxes are lemmas and propositions, and green boxes are theorems. Arrows denote when an assumption or a result is used in another proof. A use already implied by a chain of arrows is not drawn, unless it plays a separate role (labelled arrows). The proofs are grouped vertically into the three groups of proofs in Section A, contributing the finite-sample guarantee at the bottom.

Thus, ζ_n is non-decreasing in n, and given the limit $\zeta = \lim_{n\to\infty} \zeta_n$, define the process $X_s := X_s^{(n)}$ for $s < \zeta_n$. This defines a unique strong solution on $[0, \zeta)$.

(b) Non-explosion via Khasminskii's test (Khasminskii, 2012; see also Mao, 2007). Set $\Phi(x) := 1 + \|x - c\|^2$. We work conditionally on $\mathcal{F}_0$, under which Y and the constants below are fixed while W remains a Brownian motion. By Dynkin's formula applied to $\Phi(X_{s\wedge\zeta_n})$:

$$\mathbb{E}[\Phi(X_{s\wedge\zeta_n}) \mid \mathcal{F}_0] = \Phi(X_0) + \mathbb{E}\Big[\int_0^{s\wedge\zeta_n} \mathcal{L}_u\Phi(X_u)\, du \;\Big|\; \mathcal{F}_0\Big]$$

where $\mathcal{L}_u\Phi(x) := 2\langle x - c, b(u, x)\rangle + \mathrm{tr}(\Sigma\Sigma^\top)$ is the (time-dependent) generator of the equation with frozen delayed arguments.

Let $\rho_Y^2 := \sup_{s\in[0,\tau_{\min}]} \sum_{a=1}^q \|X_{s-\tau_a} - c\|^2 < \infty$, which is $\mathcal{F}_0$-measurable since K_Y is compact, so that $\|(x_0, Y_s) - c^{(q)}\|^2 \le \|x_0 - c\|^2 + \rho_Y^2$. By Assumption 2 (ii), whenever $\|(x_0, Y_s) - c^{(q)}\| \ge R$ we have $2\langle x_0 - c, G(x_0, Y_s)\rangle \le 2\beta(1 + \|x_0 - c\|^2 + \rho_Y^2) \le 2\beta(1 + \rho_Y^2)\Phi(x_0)$. Otherwise $\|x_0 - c\| < R$. For $\|x_0 - c\| \le R$, set

$$M_R := \sup\{\|G(x_0, y)\| : \|x_0 - c\| \le R,\ y \in K_Y\} < \infty \quad \text{a.s.},$$

finite by continuity of G on the compact $\bar{B}_R(c) \times K_Y$, so that $2\langle x_0 - c, G(x_0, Y_s)\rangle \le 2RM_R \le 2C_R\Phi(x_0)$ with $C_R := RM_R$, using $\Phi \ge 1$. Hence, for all $x \in \mathbb{R}^D$ and $s \in [0, \tau_{\min}]$,

$$\mathcal{L}_s\Phi(x) = 2\langle x - c, G(x, Y_s)\rangle + \mathrm{tr}(\Sigma\Sigma^\top) \le C\,\Phi(x),$$

where $C := 2\max\big(\beta(1 + \rho_Y^2), C_R\big) + \mathrm{tr}(\Sigma\Sigma^\top)$ is $\mathcal{F}_0$-measurable and a.s. finite.

As $\int_0^{s\wedge\zeta_n} \mathcal{L}_u\Phi(X_u)\, du \le C\int_0^s \Phi(X_{u\wedge\zeta_n})\, du$, Grönwall's inequality gives

$$\mathbb{E}[\Phi(X_{s\wedge\zeta_n}) \mid \mathcal{F}_0] \le \Phi(X_0)e^{Cs}, \quad s \in [0, \tau_{\min}].$$

Since $\Phi(X_{\zeta_n}) \ge 1 + n^2$ on $\{\zeta_n \le \tau_{\min}\}$, we have that

$$(1 + n^2)\mathbb{P}(\zeta_n \le \tau_{\min} \mid \mathcal{F}_0) \le \mathbb{E}[\Phi(X_{\zeta_n})\mathbf{1}_{\zeta_n\le\tau_{\min}} \mid \mathcal{F}_0] \le \Phi(X_0)e^{C\tau_{\min}}.$$

Since $\{\zeta \le \tau_{\min}\} \subseteq \{\zeta_n \le \tau_{\min}\}$, letting $n \to \infty$ yields $\mathbb{P}(\zeta \le \tau_{\min} \mid \mathcal{F}_0) = 0$ a.s., and integrating these conditional probabilities (which are bounded by one) gives $\mathbb{P}(\zeta \le \tau_{\min}) = 0$, i.e. the solution does not explode on $[0, \tau_{\min}]$. No integrability of $\Phi(X_0)e^{C\tau_{\min}}$ is needed.

Step 2: Induction on $[k\tau_{\min}, (k + 1)\tau_{\min}]$. The delayed arguments now satisfy $s - \tau_a \le k\tau_{\min}$, hence lie on the already-constructed path; parts (a)–(b) apply verbatim with $\mathcal{F}_0$ replaced by $\mathcal{F}_{k\tau_{\min}}$.

Induction over k yields a unique strong solution on $[0, \infty)$ with $\sup_{s\le T}\|X_s\| < \infty$ a.s. for every finite T.

Note: When $q = 0$ there are no delayed arguments: Step 1 applies directly on any finite interval $[0, T]$ (with Y void and $\tau_{\min}$ replaced by T), and no induction is needed. □

A.3 Population identifiability

We prove population identifiability in two stages. First, we show that the drift and the mask are pinned down on the support of the joint occupation measure: matching the risk forces $\widetilde{G} = G$ there by continuity, and faithfulness upgrades this to equality of the masks. Second, Lemma 1 shows that, under non-degenerate noise, the occupation measure has full support, so "on the support" is in fact "everywhere" and the faithfulness hypothesis reduces to plain faithfulness.

The proof follows the structure of the assumptions: a zero risk forces $\widetilde{G} = G$ on the support of $\bar{\mu}_T^q = \mathbb{R}^{D(q+1)}$ which fills the whole domain from the non-degenerate noise (Lemma 1), faithfulness prevents the learned mask from dropping a true edge, and sparsity prevents it from adding a spurious one. We compare these assumptions to those of Bellot et al. (2022) in Section B.

Theorem 1 (Identifiability)**.** *Under Assumptions 1–5, G and M are identifiable:*

$$\widetilde{G} = G \quad \textit{on } \mathrm{supp}(\bar{\mu}_T^q), \qquad \tilde{M} = M. \tag{19}$$

Moreover, under the non-degenerate noise of Assumption 1 the occupation measure has full support, $\mathrm{supp}(\bar{\mu}_T^q) = \mathbb{R}^{D(q+1)}$ by Lemma 1 (see below), so the drift equality extends to all of $\mathbb{R}^{D(q+1)}$.

Proof. Step 1: Drift identifiability on $\mathrm{supp}(\bar{\mu}_T^q)$. The true pair (G, M) is feasible by Assumption 3(i) and has risk zero. Nonnegativity and optimality (Assumption 3(ii)) therefore give $\mathcal{R}_T(\widetilde{G}) = 0$, hence by Equation 4 $\widetilde{G} = G$ $\bar{\mu}_T^q$-a.e.; G is continuous by Assumption 2 and $\widetilde{G}$ by Assumption 3(i), and the complement of a $\bar{\mu}_T^q$-null set meets every nonempty relatively open subset of $\mathrm{supp}(\bar{\mu}_T^q)$, so equality on this dense subset of the (closed) support upgrades to equality on all of $\mathrm{supp}(\bar{\mu}_T^q)$.

Step 2: Mask identifiability on the support. If $M_{ij}^a = 1$ but $\tilde{M}_{ij}^a = 0$ — that is, if a true edge were dropped — then $\widetilde{G}_j$ is constant in coordinate i of block a on all of $\mathbb{R}^{D(q+1)}$ (its mask removes that input), so $G_j = \widetilde{G}_j$ on $\mathrm{supp}(\bar{\mu}_T^q)$ takes equal values at the axis-aligned witnessing pair $z, z' \in \mathrm{supp}(\bar{\mu}_T^q)$ of Assumption 4, which is a contradiction. Hence $\tilde{M}_{ij}^a \geq M_{ij}^a$ entry-wise for all a, i.e. $\sum_a \|\tilde{M}^a\|_0 \geq \sum_a \|M^a\|_0$ with equality iff $\tilde{M}^a = M^a$ for all a; Assumption 5 forces equality. □

It remains to remove the "on the support" qualifier. The next lemma shows that non-degenerate noise makes the joint law of the lagged state charge every open set, so its support is the whole space and the identity is global.

Lemma 1 (Full support of the joint occupation measure)**.** *Under Assumptions 1 and 2, for every* $t > \tau_{\max}$ *the law of* $\big(X_t, X_{t-\tau_1}, \ldots, X_{t-\tau_q}\big)$ *on* $\mathbb{R}^{D(q+1)}$ *has full support; consequently* $\mathrm{supp}(\bar{\mu}_T^q) = \mathbb{R}^{D(q+1)}$ *for every* $T > \tau_{\max}$.

Proof. We first prove the statement for a deterministic initial history and then integrate.

Step 1: Deterministic initial segment and pilot path. Fix a deterministic initial segment $\varphi \in \mathcal{C}$, a time $t > \tau_{\max}$, a target $z = (z_0, z_1, \ldots, z_q) \in \mathbb{R}^{D(q+1)}$ and $\varepsilon > 0$, and write $\mathbb{P}_\varphi$ for the law of the solution started from $X_0^{\mathrm{seg}} = \varphi$. We must show:

$$\mathbb{P}_\varphi\left(\|X_t - z_0\| < \varepsilon, \|X_{t-\tau_1} - z_1\| < \varepsilon, \ldots, \|X_{t-\tau_q} - z_q\| < \varepsilon\right) > 0.$$

With the convention $\tau_0 = 0$, the times $t - \tau_a$ for $a = 0, 1, \ldots, q$ are distinct and strictly positive, since $t > \tau_{\max} = \tau_q$. Choose a path $h : [-\tau_{\max}, t] \to \mathbb{R}^D$ which is continuously differentiable on $[0, t]$ and satisfies

$$h|_{[-\tau_{\max}, 0]} = \varphi, \qquad h(t - \tau_a) = z_a \quad (0 \leq a \leq q).$$

Such a path exists because we are prescribing finitely many values at distinct times of $(0, t]$, together with the value $h(0) = \varphi(0)$ at the left endpoint (e.g. by spline interpolation). Note that h need not be differentiable at 0 and no regularity of φ beyond continuity is required.

Step 2: The driving control. Since Σ is invertible by Assumption 1, we may define

$$w_h(s) \;:=\; \Sigma^{-1}\left[h(s) - \varphi(0) - \int_0^s G\big(h(u), h(u-\tau_1), \ldots, h(u-\tau_q)\big)\, du\right], \qquad s \in [0, t].$$

The integrand is continuous on $[0, t]$, so w_h is continuously differentiable with $w_h(0) = 0$. In particular, w_h has square-integrable derivative and belongs to the Cameron–Martin space of the Wiener measure on $C([0, t]; \mathbb{R}^D)$ (see Nualart, 2006, Ch. 1). By construction h solves the integral equation associated with Equation 1 on $[0, t]$ driven by w_h in place of W:

$$h(s) = \varphi(0) + \int_0^s G\big(h(u), h(u-\tau_1), \ldots, h(u-\tau_q)\big)\, du + \Sigma\, w_h(s). \tag{20}$$

Step 3: Localized Grönwall estimate. Let

$$K := \Big\{(y_0, \ldots, y_q) \in \mathbb{R}^{D(q+1)} \;:\; \exists\, u \in [0, t] \text{ with } \max_{0 \leq a \leq q} \|y_a - h(u - \tau_a)\| \leq 1\Big\},$$

a compact subset of $\mathbb{R}^{D(q+1)}$, and let $L < \infty$ be a Lipschitz constant for G on K, which exists by Assumption 2 (i). Both K and L are deterministic as functions of h alone. Choose $\eta > 0$ such that

$$\|\Sigma\|_{\mathrm{op}}\, \eta\, e^{L\sqrt{q+1}\, t} \;<\; \min(1, \varepsilon), \tag{21}$$

and define the Brownian tube event and the exit time

$$A_\eta := \Big\{\sup_{s\le t}\|W_s - w_h(s)\| < \eta\Big\}, \qquad \sigma := t \wedge \inf\{s \ge 0 : \|X_s - h(s)\| \ge 1\}.$$

Since X and h agree with φ on $[-\tau_{\max}, 0]$, we have $X_u - h(u) = 0$ for every $u \in [-\tau_{\max}, 0]$, so that for $s \in [0, \sigma]$

$$\|X_{s-\tau_a} - h(s-\tau_a)\| \;\le\; D(s) := \sup_{0\le u\le s}\|X_u - h(u)\|, \qquad 0 \le a \le q.$$

Subtracting Equation 20 from the integral form of Equation 1 and using that both argument vectors lie in K before time σ, we obtain on A_η, for every $s \le \sigma$,

$$D(s) \;\le\; \|\Sigma\|_{\mathrm{op}} \sup_{u\le t}\|W_u - w_h(u)\| + L\sqrt{q+1}\int_0^s D(u)\,du \;\le\; \|\Sigma\|_{\mathrm{op}}\,\eta + L\sqrt{q+1}\int_0^s D(u)\,du,$$

the factor $\sqrt{q+1}$ coming from bounding the Euclidean norm on $\mathbb{R}^{D(q+1)}$ by $\sqrt{q+1}$ times the maximum of the blockwise norms. Grönwall's inequality and Equation 21 then give

$$D(\sigma) \;\le\; \|\Sigma\|_{\mathrm{op}}\,\eta\, e^{L\sqrt{q+1}\,t} \;<\; \min(1, \varepsilon).$$

In particular $\|X_\sigma - h(\sigma)\| < 1$ strictly, so by continuity of $s \mapsto X_s - h(s)$ the exit time cannot occur before t, i.e. $\sigma = t$ on A_η, and the estimate $\sup_{s\le t}\|X_s - h(s)\| < \varepsilon$ holds on all of $[0, t]$.

Step 4: Positivity of the Brownian tube probability. The path w_h is continuously differentiable with $w_h(0) = 0$, hence lies in the Cameron–Martin space, so the law of $W - w_h$ on $C([0,t];\mathbb{R}^D)$ is equivalent to the Wiener measure. Since the Brownian small-ball probability $\mathbb{P}(\sup_{s\le t}\|W_s\| < \eta)$ is strictly positive, we conclude that $\mathbb{P}(A_\eta) > 0$ (Stroock & Varadhan, 1972, Sec. 3). Combining with Step 3 and $h(t - \tau_a) = z_a$,

$$\mathbb{P}_\varphi\Big(\max_{0\le a\le q}\|X_{t-\tau_a} - z_a\| < \varepsilon\Big) \;\ge\; \mathbb{P}(A_\eta) \;>\; 0.$$

Step 5: Random initial history and conclusion. By Proposition 1, the solution is a measurable functional of (φ, W), so $\varphi \mapsto \mathbb{P}_\varphi(\cdot)$ is a probability kernel; using Assumption 1, the initial segment X_0^{seg} is independent of W. Conditioning on $X_0^{\mathrm{seg}} = \varphi$ and integrating the strictly positive conditional probabilities of Step 4 against $\mathrm{Law}(X_0^{\mathrm{seg}})$ yields

$$\mathbb{P}\Big(\max_{0\le a\le q}\|X_{t-\tau_a} - z_a\| < \varepsilon\Big) \;>\; 0.$$

Since $z \in \mathbb{R}^{D(q+1)}$ and $\varepsilon > 0$ were arbitrary, the law of $(X_t, X_{t-\tau_1}, \dots, X_{t-\tau_q})$ charges every nonempty open subset of $\mathbb{R}^{D(q+1)}$, i.e. it has full support. Finally, for every nonempty open $O \subset \mathbb{R}^{D(q+1)}$ the map $t \mapsto \mathbb{P}\big((X_t, X_{t-\tau_1}, \dots, X_{t-\tau_q}) \in O\big)$ is strictly positive on $(\tau_{\max}, T]$, so its integral over that interval is strictly positive. The single endpoint $t = \tau_{\max}$, at which the lagged coordinate $X_{t-\tau_q} = X_0$ is still determined by the initial history, has zero Lebesgue measure and does not affect the integral. Hence $\mathrm{supp}(\bar\mu_T^q) = \mathbb{R}^{D(q+1)}$. □

A.4 Identifiability with the penalized objective

The penalized objective decouples across output coordinates, so it is enough to argue one coordinate j at a time. For each j, we compare the true active set S_j against a candidate $\tilde S_j$ in two exhaustive cases; the minimum-signal condition $s_{\min} > 0$ makes dropping a true input strictly costly, and this yields the sufficient λ-window Equation 22.

Theorem 2 (Identifiability with the penalized objective)**.** *Let Assumptions 1, 2 and 3(i) hold. Suppose $M \neq 0$ and $s_{\min} > 0$ i.e. the system has at least one true driver, and every true driver has a measurable impact on the dynamics. This replaces Assumption 4 (faithfulness). If*

$$0 \;<\; \lambda \;<\; \frac{(T - \tau_{\max})s_{\min}^2}{d_{\max}}, \tag{22}$$

then every minimizer of J_λ over $\mathcal{H}^q$ satisfies $\tilde M = M$ and $\widetilde G = G$ on $\mathbb{R}^{D(q+1)}$. Moreover, every feasible pair with $\tilde M \neq M$ satisfies

$$J_\lambda(\tilde G, \tilde M) - J_\lambda(G, M) \ge \gamma(\lambda) := \min(\lambda, (T - \tau_{\max})s_{\min}^2 - \lambda d_{\max}) > 0$$

Proof. The true pair (G, M) is feasible by Assumption 3(i) and satisfies $\mathcal{R}_T(G) = 0$, so $J_\lambda(G, M) = \lambda \sum_{j=1}^{D} |S_j|$, where $S_j := \{(i, a) : M^a_{ij} = 1\}$ is the true active set of coordinate j and $|S_j| = \sum_{a=0}^{q} \|m^a_j\|_0 \le d_{\max}$. Let $(\widetilde{G}, \tilde{M}) \in \mathcal{H}^q$ be any feasible pair and write $\tilde{S}_j := \{(i, a) : \tilde{M}^a_{ij} = 1\}$. Since the squared $L_2(\bar{\mu}^q_T)$ risk and the ℓ_0 penalty of Equation 9 are both sums over the output coordinates, the objective gap decomposes as

$$J_\lambda(\widetilde{G}, \tilde{M}) - J_\lambda(G, M) = \sum_{j=1}^{D} \Delta_j, \qquad \Delta_j := (T - \tau_{\max}) \|G_j - \widetilde{G}_j\|^2_{L_2(\bar{\mu}^q_T)} + \lambda\big(|\tilde{S}_j| - |S_j|\big). \tag{23}$$

This decomposes the *value* of the gap at a fixed feasible pair. It therefore suffices to bound Δ_j from below in the two exhaustive cases and to sum.

Consider first a coordinate whose mask keeps every true input, $\tilde{S}_j \supseteq S_j$. Then $|\tilde{S}_j| \ge |S_j|$, and since the risk term is nonnegative we obtain $\Delta_j \ge 0$; if the containment is strict then $|\tilde{S}_j| \ge |S_j| + 1$ and

$$\Delta_j \ge \lambda > 0. \tag{24}$$

Consider next a coordinate whose mask omits some true input $(i, a) \in S_j \setminus \tilde{S}_j$, that is $\tilde{M}^a_{ij} = 0$ while $M^a_{ij} = 1$. By the masked parametrization of Assumption 3(i), the candidate $\widetilde{G}_j = \widetilde{f}_j(\tilde{m}^0_j \odot x_0, \dots, \tilde{m}^q_j \odot x_q)$ does not depend on coordinate i of block a, so it belongs to the closed subspace onto which $\Pi_{-(i,a)}$ projects (the $L_2(\bar{\mu}^q_T)$ space of the σ-algebra generated by the remaining coordinates; note that $G_j \in L_2(\bar{\mu}^q_T)$ by the finite-energy clause of Assumption 1, hence $\widetilde{G}_j \in L_2(\bar{\mu}^q_T)$ by finite risk).

Since $\Pi_{-(i,a)} G_j$ is the nearest point of that subspace to G_j, the signal strength Equation 7 lower-bounds the distance to any of its elements, and in particular

$$\|G_j - \widetilde{G}_j\|_{L_2(\bar{\mu}^q_T)} \ge \big\| G_j - \Pi_{-(i,a)} G_j \big\|_{L_2(\bar{\mu}^q_T)} = s^a_{ij} \ge s_{\min}.$$

Bounding the penalty term by $|\tilde{S}_j| \ge 0$ and $|S_j| \le d_{\max}$ gives

$$\Delta_j \ge (T - \tau_{\max}) (s^a_{ij})^2 - \lambda |S_j| \ge (T - \tau_{\max}) s^2_{\min} - \lambda d_{\max}, \tag{25}$$

which is strictly positive under the window Equation 22.

Now let $(\widetilde{G}, \tilde{M})$ be feasible with $\tilde{M} \ne M$ and let j be a coordinate in which the two masks differ. That coordinate falls into exactly one of the two cases: either it omits a true input, and Equation 25 applies, or its mask strictly contains S_j, and Equation 24 applies. In both cases $\Delta_j \ge \gamma(\lambda)$, while every remaining coordinate satisfies $\Delta_j \ge 0$ under Equation 22. Summing in Equation 23 yields $J_\lambda(\widetilde{G}, \tilde{M}) - J_\lambda(G, M) \ge \gamma(\lambda) > 0$, so no pair with an incorrect mask can be a minimizer and every minimizer satisfies $\tilde{M} = M$. Given $\tilde{M} = M$ the true pair is feasible with objective $\lambda \sum_j |S_j|$, so any minimizer has $\mathcal{R}_T(\widetilde{G}) = 0$, that is $\widetilde{G} = G$ $\bar{\mu}^q_T$-almost everywhere. G is continuous by Assumption 2 (i) and $\widetilde{G}$ by Assumption 3(i), and $\mathrm{supp}(\bar{\mu}^q_T) = \mathbb{R}^{D(q+1)}$ by Lemma 1, so $\widetilde{G} = G$ everywhere, exactly as in the proof of Theorem 1. □

A.5 Uniform moment bounds

In order to prove some ergodicity results in Section A.6, we will need some estimates on the moments of the process both point-wise and in the segment space. This section provides: first a bound on the one-point moments $\mathbb{E}\|X_t - c\|^{2p}$, obtained from Itô's formula and a Halanay comparison to handle the delayed feedback, second a lifting of these to the segment norm via the Burkholder–Davis–Gundy inequality (Karatzas & Shreve, 1991).

We first give a small lemma combining the radial bound and the residual bound from the Assumption 6 to get a bound on the whole space.

Lemma 2 (Global dissipativity estimate)**.** *Under Assumption 6, set $b_a := \beta_a + \varepsilon_a$ for $a = 1, \dots, q$, so that $B = \sum_{a=1}^{q} b_a < \beta_0$ by Equation 13. Combining the radial bound outside $\bar{B}_R(c)$ with the*

residual bound inside it, we obtain, for all $(x_0, x_1, \dots, x_q) \in \mathbb{R}^{D(q+1)}$,

$$\langle x_0 - c,\, G(x_0, x_1, \dots, x_q)\rangle \;\le\; A - \beta_0\|x_0 - c\|^2 + \sum_{a=1}^{q} b_a\|x_a - c\|^2, \tag{26}$$

where $A := \max\{\alpha,\; C_R + \beta_0 R^2\}$ *and* C_R *is the constant defined in Equation 12.*

Proof. Outside the ball, Equation 26 follows from the radial bound Equation 11, since $A \ge \alpha$ and $b_a \ge \beta_a$. Inside it, $\beta_0\|x_0 - c\|^2 \le \beta_0 R^2$, so adding and subtracting this term in the residual bound Equation 12 and using $b_a \ge \varepsilon_a$ gives Equation 26 with the constant $C_R + \beta_0 R^2 \le A$. □

Lemma 3 (One-point moment bounds)**.** *Suppose that Assumptions 1, 2 and 6 hold. For every integer* $p \ge 1$ *with* $\mathbb{E}\|X_0^{\mathrm{seg}} - c\|_{\mathcal{C}}^{2p} < \infty$ *there are constants* $\gamma_p > 0$ *and* $K_p^\star < \infty$*, depending only on* Σ*, p, the delays and the constants from the assumption bounds, such that*

$$\mathbb{E}\|X_t - c\|^{2p} \le \mathbb{E}\|X_0^{\mathrm{seg}} - c\|_{\mathcal{C}}^{2p}\, e^{-\gamma_p t} + K_p^\star \tag{27}$$

Proof. Write $Q := \Sigma\Sigma^\top$, $\Phi_1(x) := \|x - c\|^2$ and $\Phi_p := \Phi_1^{\,p}$, so that $\Phi_p(X_t) = \|X_t - c\|^{2p}$, and recall $b_a = \beta_a + \varepsilon_a$ and $B = \sum_{a=1}^q b_a$ from Lemma 2, and set $M_0 := \mathbb{E}\|X_0^{\mathrm{seg}} - c\|_{\mathcal{C}}^{2p} < \infty$.

Step 1: Generator bound for Φ_p. Since $\nabla\Phi_p = 2p\,\Phi_1^{p-1}(x - c)$ and $D^2\Phi_p = 2p\,\Phi_1^{p-1} I + 4p(p-1)\,\Phi_1^{p-2}(x-c)(x-c)^\top$, Itô's formula applied to Equation 1 gives

$$\mathcal{L}\Phi_p = 2p\,\Phi_1^{p-1}(x_0)\,\langle x_0 - c, G\rangle + p\,\Phi_1^{p-1}(x_0)\,\mathrm{tr}(Q) + 2p(p-1)\,\Phi_1^{p-2}(x_0)\big\|\Sigma^\top(x_0 - c)\big\|^2.$$

Note that the last term vanishes when $p = 1$. Bounding $\|\Sigma^\top(x_0 - c)\|^2 \le \|\Sigma\|_{\mathrm{op}}^2\,\Phi_1(x_0)$ and inserting Equation 26 yields,

$$\mathcal{L}\Phi_p \;\le\; -2p\beta_0\,\Phi_p(x_0) + 2p\sum_{a=1}^{q} b_a\,\Phi_1^{p-1}(x_0)\Phi_1(x_a) + k_p\,\Phi_1^{p-1}(x_0), \tag{28}$$

with $k_p := 2pA + p\,\mathrm{tr}(Q) + 2p(p-1)\|\Sigma\|_{\mathrm{op}}^2$. The delayed terms $\Phi_1(x_a)$ are the reason a plain Grönwall argument does not apply and a delay comparison is needed using Halanay's inequality (Halanay, 1966).

Step 2: Scaled Young inequality and the dissipativity margin. For $p > 1$, Young's inequality with conjugate exponents $p/(p-1)$ and p gives

$$\Phi_1^{p-1}(x_0)\,\Phi_1(x_a) \;\le\; \tfrac{p-1}{p}\,\Phi_p(x_0) + \tfrac{1}{p}\,\Phi_p(x_a),$$

so the cross terms in Equation 28 are bounded by $2(p-1)B\,\Phi_p(x_0) + 2\sum_a b_a\,\Phi_p(x_a)$.

The lower-order term $k_p\,\Phi_1^{p-1}(x_0)$ is treated separately using a scale η_p. Fix any

$$\eta_p \in \big(0,\, 2p(\beta_0 - B)\big), \qquad C_p := \sup_{y \ge 0}\big(k_p\,y^{p-1} - \eta_p\,y^p\big) = \frac{k_p^{\,p}\,(p-1)^{p-1}}{p^{\,p}\,\eta_p^{\,p-1}} < \infty,$$

so that $k_p\,\Phi_1^{p-1}(x_0) \le \eta_p\,\Phi_p(x_0) + C_p$; for $p = 1$ this term is already the constant k_1 and we set $\eta_1 := 0$, $C_1 := k_1$.

Substituting both bounds into Equation 28 yields the delay-dissipative estimate

$$\mathcal{L}\Phi_p \;\le\; -a_p\,\Phi_p(x_0) + 2\sum_{a=1}^{q} b_a\,\Phi_p(x_a) + C_p, \qquad a_p := 2p\beta_0 - 2(p-1)B - \eta_p. \tag{29}$$

The role of the free parameter is that the margin required by the comparison step is now automatic:

$$a_p - 2B \;=\; 2p(\beta_0 - B) - \eta_p \;>\; 0 \qquad \text{for every } p \ge 1,$$

by the choice of η_p. Applying an unscaled Young inequality to the lower-order term would instead add k_p to the coefficient of $\Phi_p(x_0)$ and can make a_p negative, which is what must be avoided when using Halanay inequality.

Step 3: A priori integrability. Before applying the inequality, we must show that $m_p(t) := \mathbb{E}\Phi_p(X_t)$ is finite and locally absolutely continuous. Let $\zeta_n := \inf\{s \geq 0 : \|X_s - c\| \geq n\}$, which increases to $+\infty$ a.s. by Proposition 1, and set

$$S_n(t) := \max\Big\{\|X_0^{\mathrm{seg}} - c\|_{\mathcal{C}}^{2p},\ \sup_{u\leq t} \Phi_p(X_{u\wedge\zeta_n})\Big\},$$

so that $\mathbb{E}S_n(t) \leq M_0 + n^{2p} < \infty$ and $S_n(s)$ dominates every delayed term $\Phi_p(X_{s-\tau_a})$ for $s \leq t$ — those with $s < \tau_a$ being read off the initial segment.

Itô's formula for $\Phi_p(X_{u\wedge\zeta_n})$ together with Equation 29, after discarding the nonpositive term $-a_p\Phi_p$, gives the following estimate

$$\mathbb{E}S_n(t) \ \leq\ 2M_0 + C_p\, t + 2B \int_0^t \mathbb{E}S_n(s)\, ds + \mathbb{E}\sup_{u\leq t}\big|N_{u\wedge\zeta_n}\big|.$$

where we wrote $N_t := \int_0^t 2p\, \Phi_1^{p-1}(X_s)(X_s - c)^\top \Sigma\, dW_s$ for the martingale part.

Since $d[N]_s \leq 4p^2\|\Sigma\|_{\mathrm{op}}^2\, \Phi_1^{2p-1}(X_s)\, ds$, the Burkholder–Davis–Gundy inequality, the factorisation $\Phi_1^{2p-1} = \Phi_p \cdot \Phi_1^{p-1}$ and Young's inequality $\sqrt{uv} \leq \frac{1}{4}u + v$ bound the last term by $\frac{1}{2}\mathbb{E}S_n(t) + C\int_0^t(1 + \mathbb{E}S_n(s))\, ds$, with C depending only on p and $\|\Sigma\|_{\mathrm{op}}$. Absorbing $\frac{1}{2}\mathbb{E}S_n(t)$ on the left and applying Grönwall's lemma gives a bound uniform in n on every finite horizon; letting $n \to \infty$ and using Fatou's lemma,

$$\mathbb{E}\sup_{u\leq H}\|X_u - c\|^{2p} < \infty \qquad \text{for every finite } H.$$

Consequently, by the same Burkholder–Davis–Gundy bound, $\mathbb{E}\sup_{u\leq H}|N_u| < \infty$, so N is a true martingale on finite horizons and expectations may be taken in Itô's formula without stopping; the positive part of $\mathcal{L}\Phi_p$ is integrable by Equation 29 and Itô's identity then controls its negative part, so m_p is locally absolutely continuous and, for almost every $t > 0$,

$$m_p'(t) \ \leq\ -a_p\, m_p(t) + 2\sum_{a=1}^{q} b_a\, m_p(t - \tau_a) + C_p. \tag{30}$$

Step 4: Halanay comparison. Consider the characteristic equation

$$\gamma_p + 2\sum_{a=1}^{q} b_a\, e^{\gamma_p \tau_a} \ =\ a_p. \tag{31}$$

Its left-hand side is strictly increasing in γ_p, equals $2B < a_p$ at $\gamma_p = 0$ by Step 2, and tends to infinity, so Equation 31 has a unique positive root $\gamma_p \in (0, a_p)$; when all b_a vanish it reduces to $\gamma_p = a_p$. Put $K_p^\star := C_p/(a_p - 2B)$ and

$$w(t) := M_0\, e^{-\gamma_p t} + K_p^\star, \qquad t \geq -\tau_{\max}.$$

A direct computation using Equation 31 shows that w satisfies Equation 30 with equality, and $w(t) \geq M_0 \geq m_p(t)$ on $[-\tau_{\max}, 0]$ since $e^{-\gamma_p t} \geq 1$ there. Set $u := m_p - w$, so that $u \leq 0$ on $[-\tau_{\max}, 0]$ and $u'(t) \leq -a_p u(t) + 2\sum_a b_a\, u(t - \tau_a)$ almost everywhere. On $[0, \tau_{\min}]$ all delayed values $u(t - \tau_a)$ are nonpositive, hence $(e^{a_p t}u)' \leq 0$ and $u(t) \leq u(0)e^{-a_p t} \leq 0$; since $\tau_{\min}$ is the smallest delay, induction over the intervals $[k\tau_{\min}, (k+1)\tau_{\min}]$ propagates $u \leq 0$ to all of $[0, \infty)$ (for $q = 0$ ordinary Grönwall suffices). Therefore $m_p(t) \leq w(t)$ for every $t \geq 0$, that is

$$\mathbb{E}\|X_t - c\|^{2p} \ \leq\ \mathbb{E}\|X_0^{\mathrm{seg}} - c\|_{\mathcal{C}}^{2p}\, e^{-\gamma_p t} + K_p^\star, \qquad t \geq 0,$$

which is the announced bound, with γ_p the positive root of Equation 31 and $K_p^\star = C_p/(a_p - 2B)$. □

With the one-point bound in hand, we lift to the segment norm and record the stationary and finite-horizon consequences (parts (b) and (c) of the lemma).

Lemma 4 (Segment moment bounds)**.** *Under the assumptions of Lemma 3 and for the same range of p, write $U_p(\varphi) := \|\varphi - c\|_{\mathcal{C}}^{2p}$ for $\varphi \in \mathcal{C}$ and let $(P_t)_{t\geq 0}$ denote the semigroup of the segment process on $\mathcal{C}$.*

(a) There are constants $A_p^{\mathrm{seg}}, B_p^{\mathrm{seg}} < \infty$ and $\gamma_p > 0$ such that

$$P_t U_p(\phi) = \mathbb{E}_\phi \|X_t^{\mathrm{seg}} - c\|_{\mathcal{C}}^{2p} \;\leq\; A_p^{\mathrm{seg}}\, e^{-\gamma_p t}\, U_p(\phi) + B_p^{\mathrm{seg}}, \qquad t \geq 0,\ \phi \in \mathcal{C}. \tag{32}$$

In particular $M_{2p}(\phi) := \sup_{t\geq 0} \mathbb{E}_\phi \|X_t^{\mathrm{seg}} - c\|_{\mathcal{C}}^{2p} \leq A_p^{\mathrm{seg}}\, U_p(\phi) + B_p^{\mathrm{seg}} < \infty$.

(b) If the segment process admits an invariant probability measure π^{seg} on $\mathcal{C}$, then

$$\int_{\mathcal{C}} \|\eta - c\|_{\mathcal{C}}^{2p}\, \pi^{\mathrm{seg}}(d\eta) < \infty. \tag{33}$$

(c) For every finite horizon H there is $C_{p,H} < \infty$ with

$$\mathbb{E}_\phi \sup_{t\leq H} \|X_t^{\mathrm{seg}} - c\|_{\mathcal{C}}^{2p} \;\leq\; C_{p,H}\big(1 + U_p(\phi)\big), \qquad \phi \in \mathcal{C}, \tag{34}$$

which is uniform over any ball $B_R := \{\phi \in \mathcal{C} : \|\phi - c\|_{\mathcal{C}} \leq R\}$.

Proof. Recall $\Phi_1(x) = \|x - c\|^2$, $\Phi_p = \Phi_1^p$ and $\sigma^2 := \|\Sigma\|_{\mathrm{op}}^2$, and note that $U_p(X_t^{\mathrm{seg}}) = \sup_{u\in[t-\tau_{\max},\, t]} \Phi_p(X_u)$, so that (a) is exactly an upgrade of the one-point control of Lemma 3 to a supremum over a window of length $\tau_{\max}$. We keep the constants b_i, $B = \sum_{i=1}^q b_i$, C_p, γ_p and $K_p^\star$ of that lemma, together with its generator bound Equation 29. All suprema below are first taken up to the localising time $\theta_N := \inf\{s : \|X_s - c\| > N\}$; the resulting constants do not depend on N, and $\theta_N \uparrow \infty$ a.s. by Proposition 1, so monotone convergence removes the localisation at the end of each step. The finite-horizon integrability established in the proof of Lemma 3 guarantees that every expectation written below is finite before the limit is taken. When $q = 0$ we use the conventions $\tau_{\max} = 0$ and $\mathcal{C} = \mathbb{R}^D$: then $U_p(X_t^{\mathrm{seg}}) = \Phi_p(X_t)$, part (a) is Lemma 3 itself, and in Step 5 the windows of length $\tau_{\max}$ are replaced by windows of unit length, to which the window estimate Equation 38 of Step 2 applies verbatim.

Step 1: A one-point envelope valid on negative times. Set $m_p(s) := \mathbb{E}_\phi \Phi_p(X_s)$ for $s \geq -\tau_{\max}$. For $s \in [-\tau_{\max}, 0]$ the path is the deterministic initial segment, so $m_p(s) \leq U_p(\phi)$. Since moreover $m_p(s) \leq U_p(\phi) e^{-\gamma_p s} + K_p^\star$ for $s \geq 0$ by Lemma 3, the inequality

$$m_p(s) \;\leq\; U_p(\phi)\, e^{-\gamma_p s} + K_p^\star \tag{35}$$

holds for every $s \geq -\tau_{\max}$.

Step 2: Window estimate. Fix $0 \leq v < w$ with $w - v \leq \tau_{\max}$. In integrated form, for $u \in [v, w]$,

$$\Phi_p(X_u) = \Phi_p(X_v) + \int_v^u \mathcal{L}\Phi_p(X_s)\, ds + (N_u - N_v),$$

with $N_s := \int_0^s 2p\, \Phi_1^{p-1}(X_r)\, (X_r - c)^\top \Sigma\, dW_r$ as in the previous lemma.

Discarding the favourable term $-a_p \Phi_p \leq 0$ in Equation 29, taking the supremum over $u \in [v, w]$ and then expectations,

$$\mathbb{E} \sup_{v\leq u\leq w} \Phi_p(X_u) \;\leq\; m_p(v) + 2\sum_{i=1}^q b_i \int_v^w m_p(s - \tau_i)\, ds + C_p(w - v) + \mathbb{E} \sup_{v\leq u\leq w} \big|N_u - N_v\big|. \tag{36}$$

Re-anchored at v, the increment $u \mapsto N_u - N_v$ is a continuous local martingale started at 0, so the Burkholder–Davis–Gundy inequality at exponent 1 applies with a universal constant C_{BDG}. Using $d[N]_s \leq 4p^2\sigma^2 \Phi_1^{2p-1}(X_s)\, ds$ and the factorisation $\Phi_1^{2p-1} = \Phi_p \cdot \Phi_1^{p-1}$,

$$\mathbb{E} \sup_{v\leq u\leq w} \big|N_u - N_v\big| \;\leq\; 2p\sigma C_{\mathrm{BDG}}\, \mathbb{E}\left[\Big(\sup_{v\leq u\leq w} \Phi_p(X_u) \Big)^{1/2} \Big(\int_v^w \Phi_1^{p-1}(X_s)\, ds \Big)^{1/2} \right].$$

Young's inequality $ab \le \frac{1}{2}a^2 + \frac{1}{2}b^2$, with the constant absorbed into the second factor and $C_N := (2p\sigma C_{\mathrm{BDG}})^2$, gives

$$\mathbb{E} \sup_{v\le u\le w} \big|N_u - N_v\big| \;\le\; \tfrac{1}{2}\,\mathbb{E} \sup_{v\le u\le w} \Phi_p(X_u) + \tfrac{C_N}{2} \int_v^w \mathbb{E}\,\Phi_1^{p-1}(X_s)\,ds, \tag{37}$$

and $\Phi_1^{p-1} \le \frac{p-1}{p}\Phi_p + \frac{1}{p} \le \Phi_p + 1$ bounds the last integral by $\int_v^w (1 + m_p(s))\,ds$. Substituting Equation 37 into Equation 36 and absorbing the half-supremum into the left-hand side,

$$\mathbb{E} \sup_{v\le u\le w} \Phi_p(X_u) \;\le\; 2m_p(v) + 4\sum_{i=1}^{q} b_i \int_v^w m_p(s-\tau_i)\,ds + 2C_p(w-v) + C_N \int_v^w \big(1+m_p(s)\big)\,ds. \tag{38}$$

Step 3: Geometric drift for the segment process. Let $t \ge \tau_{\max}$ and apply Equation 38 with $v = t - \tau_{\max}$ and $w = t$. Every time argument occurring on its right-hand side lies in $[t - 2\tau_{\max},\, t]$, so Equation 35 bounds each of them by $e^{2\gamma_p\tau_{\max}}U_p(\phi)e^{-\gamma_p t} + K_p^\star$. Writing $\Lambda_p := 2 + 4B\tau_{\max} + C_N\tau_{\max}$, we obtain Equation 32 for $t \ge \tau_{\max}$ with

$$A_p^{\mathrm{seg}} := \Lambda_p\, e^{2\gamma_p\tau_{\max}}, \qquad B_p^{\mathrm{seg}} := \Lambda_p K_p^\star + 2C_p\tau_{\max} + C_N\tau_{\max}.$$

For $0 \le t < \tau_{\max}$ the window $[t - \tau_{\max},\, t]$ straddles the origin, and the initial history is not an Itô process on its negative part. We therefore split

$$U_p(X_t^{\mathrm{seg}}) = \sup_{u\in[t-\tau_{\max},\,t]} \Phi_p(X_u) \;\le\; U_p(\phi) + \sup_{0\le u\le t} \Phi_p(X_u),$$

bound the first term directly and the second by Equation 38 with $v = 0$, $w = t$, whose length is at most $\tau_{\max}$. All time arguments in the latter lie in $[-\tau_{\max}, t] \subseteq [t - 2\tau_{\max}, t]$, so the computation above bounds the expectation of the second term by $A_p^{\mathrm{seg}}e^{-\gamma_p t}U_p(\phi) + B_p^{\mathrm{seg}}$, while the first term is at most $e^{\gamma_p\tau_{\max}}e^{-\gamma_p t}U_p(\phi)$ since $e^{-\gamma_p t} \ge e^{-\gamma_p\tau_{\max}}$ on this range. Replacing A_p^{seg} by $A_p^{\mathrm{seg}} + e^{\gamma_p\tau_{\max}}$ therefore extends Equation 32 to all $t \ge 0$. Taking the supremum over $t \ge 0$ gives the uniform bound $M_{2p}(\phi) \le A_p^{\mathrm{seg}}U_p(\phi) + B_p^{\mathrm{seg}}$ announced in (a), established without any growth assumption on G.

Step 4: Stationary integrability (b). Choose $h > 0$ so that $\kappa := A_p^{\mathrm{seg}}e^{-\gamma_p h} < 1$ and write $\varrho := B_p^{\mathrm{seg}}$, so that Equation 32 reads $P_hU_p \le \kappa U_p + \varrho$ pointwise on $\mathcal{C}$. One cannot integrate this against π^{seg} and cancel the two sides, since $\int U_p\,d\pi^{\mathrm{seg}}$ is not yet known to be finite. Therefore, set $u_N := U_p \wedge N$, a bounded measurable function. Concavity of $v \mapsto v \wedge N$ and Jensen's inequality give $P_hu_N \le (P_hU_p) \wedge N \le (\kappa U_p + \varrho) \wedge N$, and invariance of π^{seg} applied to the bounded function u_N yields

$$\int_{\mathcal{C}} \Big[u_N - \big((\kappa U_p + \varrho) \wedge N\big) + \varrho\Big]\,d\pi^{\mathrm{seg}} \;\le\; \varrho.$$

The integrand is nonnegative, because $\kappa \le 1$ implies $(\kappa U_p + \varrho) \wedge N \le (U_p \wedge N) + \varrho = u_N + \varrho$, and it converges pointwise to $(1-\kappa)U_p$ as $N \to \infty$. Fatou's lemma therefore gives $(1-\kappa)\int_{\mathcal{C}} U_p\,d\pi^{\mathrm{seg}} \le \varrho < \infty$, which is Equation 33. The argument applies separately for each p and uses no moment of π^{seg} in its hypotheses.

Step 5: Finite horizon (c). Cover $[0, H]$ by the $n := \lceil H/\tau_{\max}\rceil$ windows $[k\tau_{\max}, (k+1)\tau_{\max}] \cap [0, H]$, $k = 0, \dots, n-1$, each of length at most $\tau_{\max}$. Since the supremum over a union is at most the sum of the suprema,

$$\mathbb{E}_\phi \sup_{t\le H} U_p(X_t^{\mathrm{seg}}) = \mathbb{E}_\phi \sup_{u\in[-\tau_{\max},\,H]} \Phi_p(X_u) \;\le\; U_p(\phi) + \sum_{k=0}^{n-1} \mathbb{E}_\phi \sup_{u\in[k\tau_{\max},(k+1)\tau_{\max}]\wedge H} \Phi_p(X_u),$$

and each summand is bounded by Equation 38 together with the envelope Equation 35, which gives $m_p(s) \le U_p(\phi) + K_p^\star$ for every $s \ge -\tau_{\max}$. Collecting the n contributions yields Equation 34 with a constant $C_{p,H}$ depending only on p, H, the delays and the constants of Lemma 3, and in particular uniform over $\phi \in B_R$ for each fixed R. □

A.6 ERGODICITY OF THE TRUE PROCESS

Having those moment estimates both in the segment space and point-wise at hand, it is now possible to show that the true process admits a unique invariant probability and that we have geometric ergodicity in Wasserstein-d distance for some distance on the segment space. The proof is based on the application of the generalized Harris theorem for delayed SDEs by Hairer et al. (2011).

Besides the Lyapunov condition provided by Lemma 4, Harris' theorem needs the transition laws started from any two histories in a bounded set to overlap. For SDDEs this is the tricky part: indeed, the two initial segments may differ over the whole window $[-\tau_{\max}, 0]$, while the noise only acts on the present. The idea of the proof is to steer one copy of the process onto the other with a control acting on $[0, h]$. After time $h > \tau_{\max}$ the two segments coincide, and since Σ is invertible, Girsanov's theorem shows that the price paid for this control is a change of measure with a bounded energy.

Lemma 5 (Uniform overlap of the segment transition kernels)**.** *Under Assumptions 1, 2 and 6, fix $h > \tau_{\max}$ and $R_0 > 0$, and set $B_{R_0} := \{\varphi \in \mathcal{C} : \|\varphi - c\|_{\mathcal{C}} \le R_0\}$. There exists $\delta > 0$ depending on h and R_0 (and on G through a local Lipschitz constant, see Equation 41) such that*

$$\sup_{\varphi,\psi \in B_{R_0}} \|P_h(\varphi, \cdot) - P_h(\psi, \cdot)\|_{\mathrm{TV}} \le 1 - \delta. \tag{39}$$

Proof. Fix $\varphi, \psi \in B_{R_0}$, let X start from φ, and put $t_1 := h - \tau_{\max} > 0$ and $v := \|\varphi - \psi\|_{\mathcal{C}} \le 2R_0$. Define the deterministic difference pilot path

$$g(t) := \begin{cases} \varphi(t) - \psi(t), & -\tau_{\max} \le t \le 0, \\ (1 - t/t_1)\big(\varphi(0) - \psi(0)\big), & 0 \le t \le t_1, \\ 0, & t_1 \le t \le h. \end{cases}$$

Then g is continuous, $\sup_t \|g(t)\| \le v$, and its restriction to $[0, h]$ is absolutely continuous with $\|g'(t)\| \le (v/t_1)\mathbf{1}_{(0,t_1)}(t)$ almost everywhere.

Step 1: Localized control. The finite-horizon estimate in Lemma 4 gives

$$\sup_{\varphi \in B_{R_0}} \mathbb{E}_\varphi \sup_{0 \le t \le h} \|X_t - c\|^2 < \infty.$$

Consequently, we can choose a deterministic $M > R_0$ such that, with $\sigma := \inf\{t \ge 0 : \|X_t - c\| \ge M\}$,

$$\inf_{\varphi \in B_{R_0}} \mathbb{P}_\varphi(\sigma > h) \ge \tfrac{1}{2}.$$

Write $Z_t := (X_t, X_{t-\tau_1}, \dots, X_{t-\tau_q})$ and $g_t^{\arg} := (g(t), g(t-\tau_1), \dots, g(t-\tau_q))$. Define the progressively measurable control

$$u_t := \mathbf{1}_{\{t < \sigma\}} \Sigma^{-1} \big[G(Z_t) - G(Z_t - g_t^{\arg}) - g'(t)\big], \qquad 0 \le t \le h. \tag{40}$$

All arguments in this expression belong to a fixed compact subset of $\mathbb{R}^{D(q+1)}$, uniformly over $\varphi, \psi \in B_{R_0}$ in the finite times $t < \sigma$. Let L be a Lipschitz constant of G on a ball containing this set. Since $\|g_t^{\arg}\| \le \sqrt{q+1}\, v$, the control has the deterministic energy bound

$$\int_0^h \|u_t\|^2 \, dt \le 8R_0^2 \|\Sigma^{-1}\|_{\mathrm{op}}^2 \left((q+1)L^2 h + \frac{1}{t_1}\right) =: K_{h,R_0} < \infty. \tag{41}$$

Set $Y_t = X_t - g(t)$ up to σ, and, when $\sigma < h$, continue Y after σ using the original SDDE with the same Brownian motion and its segment at σ as initial condition. Strong existence and pathwise uniqueness justify this continuation. By substitution, Y satisfies

$$dY_t = G(Y_t, Y_{t-\tau_1}, \dots, Y_{t-\tau_q})\, dt + \Sigma\, dW_t + \Sigma u_t\, dt, \qquad Y_0^{\mathrm{seg}} = \psi.$$

On the event $E := \{\sigma > h\}$, the two terminal segments agree because $g = 0$ on $[h - \tau_{\max}, h]$. Thus

$$Y_h^{\mathrm{seg}} = X_h^{\mathrm{seg}} \quad \text{on } E, \qquad \text{with } \mathbb{P}(E) \ge \tfrac{1}{2}.$$

Step 2: Correction of the controlled law. Define

$$\Lambda := \exp\left(-\int_0^h u_t \cdot dW_t - \frac{1}{2}\int_0^h \|u_t\|^2\, dt\right), \qquad d\mathbb{Q} := \Lambda\, d\mathbb{P}.$$

The deterministic bound Equation 41 implies Novikov's condition. Under $\mathbb{Q}$, the process $W_t + \int_0^t u_s\, ds$ is Brownian, so uniqueness in law gives $\mathcal{L}_{\mathbb{Q}}(Y_h^{\mathrm{seg}}) = P_h(\psi, \cdot)$. The exponential martingale with integrand u also has expectation one, and therefore

$$\mathbb{E}[\Lambda^{-1}] = \mathbb{E}\left[\exp\left(\int_0^h u_t \cdot dW_t - \frac{1}{2}\int_0^h \|u_t\|^2\, dt\right)\exp\left(\int_0^h \|u_t\|^2\, dt\right)\right] \le e^{K_{h,R_0}}.$$

Set $\ell := e^{-K_{h,R_0}}/4$. Markov's inequality gives $\mathbb{P}(\Lambda < \ell) \le 1/4$, hence

$$\mathbb{E}\big[\mathbf{1}_E \min(1, \Lambda)\big] \ge \ell\, \mathbb{P}\big(E \cap \{\Lambda \ge \ell\}\big) \ge \frac{e^{-K_{h,R_0}}}{16}.$$

For a Borel set $A \subseteq \mathcal{C}$, define

$$\nu(A) := \mathbb{E}\big[\mathbf{1}_E \min(1, \Lambda)\mathbf{1}_{\{X_h^{\mathrm{seg}} \in A\}}\big].$$

Since $X_h^{\mathrm{seg}} = Y_h^{\mathrm{seg}}$ on E, this measure is dominated by both $P_h(\varphi, \cdot)$ and $P_h(\psi, \cdot)$. Its mass is at least $e^{-K_{h,R_0}}/16$, which proves Equation 39 with $\delta := e^{-K_{h,R_0}}/16$. □

Theorem 3 (Ergodicity: existence, uniqueness, exponential convergence)**.** *Under Assumptions 1, 2 and 6, the segment process admits a unique invariant probability measure π^{seg} on $\mathcal{C}$. Write $P_t(\varphi, \cdot) := \mathcal{L}(X_t^{\mathrm{seg}} \mid X_0^{\mathrm{seg}} = \varphi)$ and $V(\varphi) := \|\varphi - c\|_{\mathcal{C}}^2$, where c also denotes the constant segment with value c. The invariant measure satisfies*

$$\int_{\mathcal{C}} \|\eta - c\|_{\mathcal{C}}^{2p}\, \pi^{\mathrm{seg}}(d\eta) < \infty, \qquad p \ge 1 \text{ an integer.}$$

There exist $C \ge 1$ and $\gamma > 0$ such that, for every $\epsilon > 0$, the metric $d(\varphi, \psi) := 1 \wedge \epsilon^{-1}\|\varphi - \psi\|_{\mathcal{C}}$ satisfies

$$\mathcal{W}_d\big(P_t(\varphi, \cdot), \pi^{\mathrm{seg}}\big) \le \|P_t(\varphi, \cdot) - \pi^{\mathrm{seg}}\|_{\mathrm{TV}} \le Ce^{-\gamma t}\big(1 + V(\varphi)\big), \qquad t \ge 0, \quad \varphi \in \mathcal{C}. \tag{15}$$

Proof. By Proposition 1, time-homogeneity and pathwise uniqueness define a Markov semigroup $(P_t)_{t\ge 0}$ on $\mathcal{C}$.

Step 1: Lyapunov estimate. For $V(\varphi) = \|\varphi - c\|_{\mathcal{C}}^2$, the decaying segment-moment bound in Lemma 4 provides constants $A_V \ge 1$, $B_V < \infty$ and $\gamma_V > 0$ such that

$$P_t V(\varphi) \le A_V e^{-\gamma_V t} V(\varphi) + B_V, \qquad t \ge 0. \tag{42}$$

Choose $h > \tau_{\max}$ sufficiently large that $A_V e^{-\gamma_V h} < 1$. Every sublevel set $\{V \le S\}$ is the supremum-norm ball $B_{\sqrt{S}}$, and Lemma 5 gives uniform transition overlap on this set for P_h. Since Lemma 5 holds for every radius R_0 at the same time h, this overlap is available on every sublevel set of V, however large; in particular, the size of the level set required in Harris' theorem (of order $B_V/(1 - A_V e^{-\gamma_V h})$) imposes no restriction.

Step 2: Harris' theorem in total variation. The estimate Equation 42 is a Lyapunov condition, and Equation 39 is precisely the pairwise total-variation small-set condition in Hairer et al. (2011, Theorem 1.5). The choice of h satisfies the time requirement $h > \gamma_V^{-1} \log A_V$ in that theorem. It follows that $(P_t)_{t\ge 0}$ has a unique invariant probability measure π^{seg} and that, for some $C \ge 1$ and $\gamma > 0$,

$$\|P_t(\varphi, \cdot) - \pi^{\mathrm{seg}}\|_{\mathrm{TV}} \le Ce^{-\gamma t}\big(1 + V(\varphi)\big), \qquad t \ge 0, \quad \varphi \in \mathcal{C}.$$

Only pairwise overlap is required here; no common minorising measure for an entire sublevel set is asserted.

Step 3: Stationary moments and Wasserstein convergence. Now that an invariant measure exists, the stationary-integrability part of Lemma 4 gives all the stated polynomial moments. For any two probability measures μ, ν on $\mathcal{C}$, a maximal coupling has probability $\|\mu - \nu\|_{\mathrm{TV}}$ of unequal coordinates. Since $d \le 1$ and $d(\varphi, \varphi) = 0$, this coupling gives $\mathcal{W}_d(\mu, \nu) \le \|\mu - \nu\|_{\mathrm{TV}}$. Combining this inequality with the preceding total-variation estimate proves Equation 15. □

Remark 1 (Stabilisation of the identifiability constants)**.** *Under Theorem 3, averaging Equation 15 over $t \in [\tau_{\max}, T]$ against the initial law and applying the evaluation map at the lags gives*

$$\|\bar{\mu}_T^q - \mu_\infty^q\|_{\mathrm{TV}} \le \frac{C\big(1 + \mathbb{E}V(X_0^{\mathrm{seg}})\big)}{\gamma(T - \tau_{\max})}\big(e^{-\gamma\tau_{\max}} - e^{-\gamma T}\big),$$

where μ_∞^q is the stationary joint law of $(X_0, X_{-\tau_1}, \dots, X_{-\tau_q})$ under π^{seg}. Total-variation convergence alone does not control squared errors of an unbounded drift (the stationary drift need not even be square-integrable). If, in addition, $\|G\|^2$ is uniformly integrable with respect to the family $\{\bar{\mu}_T^q\}_{T>\tau_{\max}} \cup \{\mu_\infty^q\}$, then the signal strengths s_{ij}^a and hence $s_{\min}$ stabilize as $T \to \infty$ to their stationary counterparts, and the λ window Equation 22 scales linearly in T with an asymptotically constant slope. This is what makes the finite-horizon constants of Theorem 2 meaningful uniformly in T, rather than per-$(T,$ initial law$)$. The finite-sample analysis of Section A.8 avoids this issue by working directly under the stationary law and imposing a finite stationary drift energy in Equation 48 explicitly.

A.7 Mixing of the sampled process

The finite-sample analysis of Section A.8 requires the sampled lagged state to forget its past at an exponential rate, in the sense of β-mixing. Since β-mixing is a total-variation property, a Wasserstein contraction of the segment process would not be enough. Theorem 3, however, already provides geometric ergodicity in total variation: despite the infinite-dimensional state space, the noise is additive and non-degenerate, so that for $h > \tau_{\max}$ the kernels $P_h(\varphi, \cdot)$ and $P_h(\psi, \cdot)$ overlap (Lemma 5). The mixing rate is therefore a corollary of Theorem 3. We first bound the β-coefficient of a stationary Markov process by the π-average distance of its kernel from equilibrium (Lemma 6), and then pass from the segment process to the lagged state and to its deterministic subsamples by monotonicity (Theorem 5). No assumption beyond those of Theorem 3 is needed, and the non-delayed case $q = 0$ is included.

Definition 1 (β-mixing coefficient)**.** *We normalize the total-variation distance of two probability measures μ, ν on a common measurable space as*

$$\|\mu - \nu\|_{\mathrm{TV}} := \sup_A |\mu(A) - \nu(A)| = \tfrac{1}{2}\int |\mathrm{d}\mu - \mathrm{d}\nu| \ \in [0, 1].$$

For sub-σ-algebras $\mathcal{A}, \mathcal{B} \subseteq \mathcal{F}$, the β-mixing (absolute regularity) coefficient *is*

$$\beta(\mathcal{A}, \mathcal{B}) := \sup \ \tfrac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J} \left|\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\,\mathbb{P}(B_j)\right|,$$

the supremum over all finite $\mathcal{A}$-measurable partitions $\{A_i\}_{i\le I}$ and $\mathcal{B}$-measurable partitions $\{B_j\}_{j\le J}$ of Ω (Bradley, 2005). Since any pair of partitions admissible for $(\mathcal{A}', \mathcal{B}')$ is admissible for $(\mathcal{A}, \mathcal{B})$,

$$\mathcal{A}' \subseteq \mathcal{A},\ \mathcal{B}' \subseteq \mathcal{B} \quad \Longrightarrow \quad \beta(\mathcal{A}', \mathcal{B}') \le \beta(\mathcal{A}, \mathcal{B}). \tag{43}$$

For a process $(Y_s)_{s\ge s_0}$, its β-mixing coefficient at gap $t > 0$ *is*

$$\beta_Y(t) := \sup_{s \ge s_0} \ \beta\Big(\sigma(Y_u :\ s_0 \le u \le s),\ \sigma(Y_u :\ u \ge s + t)\Big), \tag{44}$$

and for a sequence $(Y_k)_{k\ge 0}$ we set $\beta_Y(m) := \sup_{k\ge 0}\beta\big(\sigma(Y_j : j \le k), \sigma(Y_j : j \ge k + m)\big)$, $m \ge 1$. By Equation 43, β_Y is nonincreasing. The process is β-mixing *if $\beta_Y(t) \to 0$ as $t \to \infty$, and* exponentially β-mixing *if $\beta_Y(t) \le C_\beta\, e^{-\lambda_{\mathrm{erg}} t}$ for constants $C_\beta < \infty$, $\lambda_{\mathrm{erg}} > 0$. For a stationary process indexed by $\mathbb{R}$ with past $\sigma(Y_u : u \le s)$, the supremum is redundant and one recovers the classical coefficient $\beta(\mathcal{F}_{\le 0}, \mathcal{F}_{\ge t})$; the one-sided form Equation 44 avoids constructing a two-sided extension of the solution.*

Lemma 6 (β-coefficient of a stationary Markov process)**.** *Let $(\xi_s)_{s\ge 0}$ be a time-homogeneous Markov process on a Polish space E, Markov with respect to its natural filtration $\mathcal{G}_s := \sigma(\xi_u :$ $0 \le u \le s)$, with transition kernels $(P_t)_{t\ge 0}$ and invariant probability measure π, and started from $\xi_0 \sim \pi$. Then, for every $s \ge 0$ and $t > 0$,*

$$\beta\Big(\mathcal{G}_s,\ \sigma(\xi_u : u \ge s + t)\Big) = \int_E \|P_t(x, \cdot) - \pi\|_{\mathrm{TV}}\, \pi(\mathrm{d}x), \tag{45}$$

so that, in particular, $\beta_\xi(t)$ is given by the right-hand side.

This identity is due to Davydov (1974) for Markov chains on a general state space, and its continuous-time version is standard (see, e.g., Masuda (2007)). Only the inequality "$\leq$" is used below. It follows by conditioning each future event on $\mathcal{G}_s$: by the Markov property and stationarity, the conditional law of $(\xi_{s+t+r})_{r\geq 0}$ is the law of the process started from $P_t(\xi_s,\cdot)$, while its unconditional law is the law of the process started from π, and a Markov kernel does not increase total variation. For the reverse inequality, restrict both σ-algebras to the endpoints using Equation 43: $\beta(\sigma(\xi_s),\sigma(\xi_{s+t}))$ is the total-variation distance between the joint law $\pi(\mathrm{d}x)P_t(x,\mathrm{d}y)$ of (ξ_s,ξ_{s+t}) and the product law $\pi(\mathrm{d}x)\pi(\mathrm{d}y)$, which equals the right-hand side of Equation 45.

Theorem 5 (Exponential β-mixing of the lagged state)**.** *Let Assumptions 1, 2 and 6 hold, let $C \geq 1$ and $\gamma > 0$ be the constants of Theorem 3, and take the process in its stationary regime, $X_0^{\mathrm{seg}} \sim \pi^{\mathrm{seg}}$ independent of W. Set*

$$M_2^\pi := \int_{\mathcal{C}} \|\eta - c\|_{\mathcal{C}}^2 \, \pi^{\mathrm{seg}}(\mathrm{d}\eta) < \infty, \qquad C_\beta := C\big(1 + M_2^\pi\big),$$

and write $Z_t := (X_t, X_{t-\tau_1}, \ldots, X_{t-\tau_q})$, $t \geq 0$. Then:

- *(i) the segment process is exponentially β-mixing: $\beta_{X^{\mathrm{seg}}}(t) \leq C_\beta e^{-\gamma t}$ for all $t > 0$;*
- *(ii) the lagged state is exponentially β-mixing: $\beta_Z(t) \leq C_\beta e^{-\gamma t}$ for all $t > 0$, and the same bound holds for $(Z_t)_{t\geq \tau_{\max}}$;*
- *(iii) for every deterministic grid $0 \leq t_0 < t_1 < \cdots$, the sampled sequence $Z_k := Z_{t_k}$ satisfies, for all $k \geq 0$ and $m \geq 1$,*

$$\beta\Big(\sigma(Z_j : j \leq k),\ \sigma(Z_j : j \geq k+m)\Big) \leq C_\beta \, e^{-\gamma(t_{k+m}-t_k)}; \tag{46}$$

in particular $\beta_{(Z_k)}(m) \leq C_\beta e^{-\gamma \Delta_{\min} m}$ whenever $\min_k(t_{k+1}-t_k) \geq \Delta_{\min} > 0$.

Consequently, the physical-time mixing estimate Equation 50 used in Section A.8 is implied by Assumptions 1, 2 and 6, with $\lambda_{\mathrm{erg}} := \gamma$ and C_β as above.

Proof. Step 1: Stationary regime and Markov property. By Theorem 3 with $p = 1$, $M_2^\pi < \infty$, so the initial condition $X_0^{\mathrm{seg}} \sim \pi^{\mathrm{seg}}$ satisfies the initial condition of Assumption 1, and Proposition 1 provides a unique strong solution; invariance of π^{seg} gives $X_s^{\mathrm{seg}} \sim \pi^{\mathrm{seg}}$ for every $s \geq 0$. Finite stationary drift energy is not needed for mixing; it is imposed separately in Section A.8. As in the proof of Theorem 3, $(X_t^{\mathrm{seg}})_{t\geq 0}$ is a time-homogeneous Markov process with kernels $(P_t)_{t\geq 0}$ with respect to the filtration generated by X_0^{seg} and W. Since its natural filtration $\mathcal{G}_s := \sigma(X_u^{\mathrm{seg}} : 0 \leq u \leq s)$ is smaller and X^{seg} is adapted to it, the tower property shows that X^{seg} is also Markov with respect to $(\mathcal{G}_s)$. The space $\mathcal{C}$ is a separable Banach space, hence Polish, so Lemma 6 applies with $E = \mathcal{C}$ and $\pi = \pi^{\mathrm{seg}}$.

Step 2: Segment process, item (i). Fix $s \geq 0$ and $t > 0$. By Lemma 6 and Equation 15,

$$\begin{aligned}\beta\Big(\mathcal{G}_s,\ \sigma(X_u^{\mathrm{seg}} : u \geq s+t)\Big) &= \int_{\mathcal{C}} \|P_t(\eta,\cdot) - \pi^{\mathrm{seg}}\|_{\mathrm{TV}} \, \pi^{\mathrm{seg}}(\mathrm{d}\eta) \\ &\leq Ce^{-\gamma t} \int_{\mathcal{C}} \big(1 + \|\eta - c\|_{\mathcal{C}}^2\big) \, \pi^{\mathrm{seg}}(\mathrm{d}\eta) = C_\beta \, e^{-\gamma t}. \end{aligned} \tag{47}$$

Taking the supremum over $s \geq 0$ gives (i).

Step 3: Lagged state, item (ii). The evaluation map $\mathrm{ev} : \mathcal{C} \to \mathbb{R}^{D(q+1)}$, $\mathrm{ev}(\eta) := (\eta(0), \eta(-\tau_1), \ldots, \eta(-\tau_q))$, is continuous, and $Z_u = \mathrm{ev}(X_u^{\mathrm{seg}})$ for every $u \geq 0$. Hence

$$\sigma(Z_u : 0 \leq u \leq s) \subseteq \mathcal{G}_s, \qquad \sigma(Z_u : u \geq s+t) \subseteq \sigma(X_u^{\mathrm{seg}} : u \geq s+t),$$

and Equation 43 together with Equation 47 gives $\beta_Z(t) \leq C_\beta e^{-\gamma t}$. Restricting the index set to $u \geq \tau_{\max}$ shrinks both σ-algebras and the range of the supremum in Equation 44, so the same bound holds for $(Z_t)_{t\geq \tau_{\max}}$.

Step 4: Deterministic subsamples, item (iii). Fix $k \geq 0$, $m \geq 1$ and set $s := t_k$, $t := t_{k+m} - t_k > 0$. Then $\sigma(Z_j : j \leq k) \subseteq \mathcal{G}_s$ and $\sigma(Z_j : j \geq k+m) \subseteq \sigma(X_u^{\mathrm{seg}} : u \geq s+t)$, so Equation 46 follows from Equation 43 and Equation 47. No regularity of the spacings is used. If all spacings are at least $\Delta_{\min}$, then $t_{k+m} - t_k \geq m\Delta_{\min}$, and taking the supremum over k gives the last claim. □

Remark 2 (Scope of Theorem 5). *For $q = 0$ the segment space reduces to $\mathbb{R}^D$ ($\tau_{\max} = 0$) and Theorem 5 covers the Markovian SDE case. The additive form of the noise is what makes the total-variation proof valid: if the diffusion coefficient depended on the delayed state, the quadratic variation of the solution could reveal the initial segment, the kernels $P_h(\varphi, \cdot)$ and $P_h(\psi, \cdot)$ could then be mutually singular (Hairer et al., 2011), and Lemma 5 (hence the bound Equation 47) could fail.*

A.8 Finite-sample identifiability

We now move from the population results to what can be guaranteed from a single trajectory observed on a finite grid. The argument compares the empirical excess risk to its population counterpart uniformly over the model class, and concludes with the margin $\gamma(\lambda)$ of Theorem 2. We observe the stationary solution, with $X_0^{\mathrm{seg}} \sim \pi^{\mathrm{seg}}$ independent of the future Brownian motion W. Write $\mathbb{P}_\pi$ and $\mathbb{E}_\pi$ for probability and expectation under this law, including both the initial segment and the subsequent noise. For $Z_t := (X_t, X_{t-\tau_1}, \dots, X_{t-\tau_q})$, let μ_∞^q be its stationary distribution. Thus $\bar{\mu}_T^q = \mu_\infty^q$, and the signal strengths in Equation 7 do not depend on T in this section. The finite-energy condition in Assumption 1 is imposed on this stationary solution:

$$m_G^2 := \mathbb{E}_\pi \|G(Z_0)\|^2 = \int \|G(z)\|^2 \, \mu_\infty^q(dz) < \infty. \tag{48}$$

Observations are available on a deterministic grid $0 = t_0 < t_1 < \cdots < t_n = T$, with $\Delta_k := t_{k+1} - t_k$ and $\Delta_{\max} := \max_{k<n} \Delta_k \le 1$. We assume that $\tau_{\max}$ is an observation time and that $X_{t_k - \tau_a}$ is observed exactly for every used left endpoint $t_k \ge \tau_{\max}$ and every lag a. An equally spaced grid has this property when each delay is an integer multiple of its spacing. All sums over k below are over $k < n$ with $t_k \ge \tau_{\max}$, and we set

$$Z_k := Z_{t_k}, \qquad T_\circ := \sum_k \Delta_k = T - \tau_{\max}.$$

Assumption 7 (Finite-sample model class). *Let $\mathcal{H}_n^q$ denote the restricted class of deterministic models derived from Assumption 3(i) satisfying the following two conditions:*

(i) ***Local finite covering:*** *For every radius $\rho \ge 1$ and tolerance $\epsilon > 0$, we can select a finite number of representative functions from within $\mathcal{H}_n^q$ (an internal ϵ-net) such that every function in the class is within a maximum distance of ϵ from at least one representative over the compact set $K_\rho := \{z \in \mathbb{R}^{D(q+1)} : \|z - c^{(q)}\| \le \rho\}$. The minimal number of representatives required to cover the class this way is denoted by $N_n(\rho, \epsilon)$.*

(ii) ***Stationary envelope:*** *The pointwise diameter of the class, defined as $\mathcal{E}(z) := \sup_{\tilde{G}_1, \tilde{G}_2 \in \mathcal{H}_n^q} \|\tilde{G}_1(z) - \tilde{G}_2(z)\|$, has a finite second moment under the stationary measure: $m_{\mathcal{E}}^2 := \int \mathcal{E}(z)^2 \mu_\infty^q(dz) < \infty$.*

The local covering condition makes $\mathcal{E}$ locally bounded, and it is measurable as a supremum of continuous functions. Since the true drift is feasible, $\|\widetilde{G}(z) - G(z)\| \le \mathcal{E}(z)$ for every candidate. Thus the finite second moment of the previous assumption controls prediction errors, not the absolute size of a dissipative drift. Any deterministic locally bounded measurable upper bound for these errors with finite stationary second moment can be used in its place. All class-dependent quantities may depend on n.

Define the stationary mean-square modulus of the true drift by

$$\omega_G(h) := \sup_{0 \le u \le h} \left(\mathbb{E}_\pi \|G(Z_u) - G(Z_0)\|^2\right)^{1/2}, \qquad 0 \le h \le 1. \tag{49}$$

Continuity and Equation 48 imply $\omega_G(h) \to 0$ as $h \downarrow 0$, as shown in the proof below. By Theorem 5, the total-variation conclusion of Theorem 3 and the stationary moment bound give the physical-time mixing estimate

$$\beta_Z(u) \le C_\beta e^{-\lambda_{\mathrm{erg}} u}, \qquad u > 0, \tag{50}$$

where $\lambda_{\mathrm{erg}} = \gamma$ and $C_\beta = C\left(1 + \int \|\varphi - c\|_{\mathcal{C}}^2 \, \pi^{\mathrm{seg}}(d\varphi)\right)$ may be used, with C, γ from Theorem 3.

Consider the empirical criterion, and assume that a measurable global minimizer exists (a finite local cover does not by itself guarantee this):

$$\begin{aligned}\widehat{\mathcal{R}}_n(\widetilde{G}) &:= \sum_k \Delta_k \Big\| \frac{X_{t_{k+1}} - X_{t_k}}{\Delta_k} - \widetilde{G}(Z_k) \Big\|^2, \\ (\widehat{G}, \widehat{M}) &\in \operatorname*{arg\,min}_{(\widetilde{G},\tilde{M})\in\mathcal{H}_n^q} \left\{ \widehat{\mathcal{R}}_n(\widetilde{G}) + \lambda \sum_{a=0}^{q} \|\tilde{M}^a\|_0 \right\}.\end{aligned} \tag{16}$$

The relevant comparison is between the empirical excess risk $\widehat{\mathcal{R}}_n(\widetilde{G}) - \widehat{\mathcal{R}}_n(G)$ and

$$\mathcal{R}_T(\widetilde{G}) = T_\circ \|\widetilde{G} - G\|^2_{L^2(\mu_\infty^q)}.$$

Theorem 4 (Finite-sample exact support recovery, L_0 estimator)**.** *Let Assumptions 1, 2, 6 and 7 hold for the stationary observations above. Fix $\delta \in (0,1)$ and choose the deterministic quantities in Equation 56–Equation 57 below. With probability at least $1-\delta$,*

$$\sup_{\widetilde{G}\in\mathcal{H}_n^q} \Big| \big[\widehat{\mathcal{R}}_n(\widetilde{G}) - \widehat{\mathcal{R}}_n(G)\big] - \mathcal{R}_T(\widetilde{G}) \Big| \le \varepsilon_n(\delta;\rho,\epsilon), \tag{51}$$

where the explicit bound is given in Equation 58. Suppose additionally that $M \neq 0$, $s_{\min} > 0$, and

$$0 < \lambda < \frac{T_\circ s_{\min}^2}{d_{\max}}, \qquad \gamma(\lambda) := \min\{\lambda, T_\circ s_{\min}^2 - \lambda d_{\max}\} > 0. \tag{52}$$

If

$$\varepsilon_n(\delta;\rho,\epsilon) < \gamma(\lambda), \tag{53}$$

then every measurable minimizer in Equation 16 satisfies

$$\mathbb{P}_\pi\left(\widehat{M}^a = M^a \text{ for all } a = 0,\dots,q, \quad \|\widehat{G} - G\|^2_{L^2(\mu_\infty^q)} \le \frac{\varepsilon_n(\delta;\rho,\epsilon)}{T_\circ} \right) \ge 1-\delta. \tag{54}$$

The conclusion applies to both $q = 0$ and $q \ge 1$. It identifies the mask exactly, not the drift exactly from finitely many observations. With $\lambda = \lambda_0 T_\circ$ for $0 < \lambda_0 < s_{\min}^2/d_{\max}$, recovery follows whenever the displayed bound divided by $T_\circ$ is smaller than $\min\{\lambda_0, s_{\min}^2 - \lambda_0 d_{\max}\}$. Thus consistency requires a joint condition on the observation duration, mesh and class complexity; large n alone does not suffice.

For $\rho \ge 1$, define

$$E_\rho := 1 \vee \sup_{z\in K_\rho} \mathcal{E}(z), \qquad v_\rho := \int_{K_\rho^c} \mathcal{E}(z)^2\, \mu_\infty^q(dz), \qquad \sigma_{\mathrm{op}} := \|\Sigma\|_{\mathrm{op}}, \quad \sigma_F := \sqrt{\mathrm{tr}(\Sigma\Sigma^\top)}. \tag{55}$$

Here E_ρ bounds errors on the truncation region and $v_\rho \to 0$ is the remaining stationary squared-error mass. For an integer $p \ge 1$, let

$$S_p := \mathbb{E}_\pi \sup_{0\le u\le 1} \|X_u^{\mathrm{seg}} - c\|_{\mathcal{C}}^{2p} < \infty.$$

Finiteness follows by integrating the finite-horizon estimate of Lemma 4(c) against π^{seg} and using its stationary moments.

For $\delta \in (0,1)$, choose $p \ge 1$ and $\epsilon > 0$, and set

$$\eta := \delta/6, \qquad J := \max\{1, \lceil T_\circ \rceil\}, \qquad \rho \ge \max\left\{1, \left(\frac{J(q+1)^p S_p}{\eta}\right)^{1/(2p)}\right\}. \tag{56}$$

We define some more quantities: the block length and logarithmic complexity,

$$b := \max\left\{1, \frac{1}{\lambda_{\mathrm{erg}}} \log\frac{C_\beta J}{\eta}\right\}, \qquad N := N_n(\rho,\epsilon), \qquad \Lambda := \log\frac{4N}{\eta}. \tag{57}$$

A valid bound in Equation 51 is then

$$\begin{aligned}\varepsilon_n(\delta;\rho,\epsilon) = \frac{2m_{\mathcal{E}}T_\circ}{\eta}\,\omega_G(\Delta_{\max}) \\ + 2\sigma_{\mathrm{op}}E_\rho\sqrt{2T_\circ\Lambda} + \frac{2\epsilon\sigma_F}{\eta}\sum_k\sqrt{\Delta_k} \\ + E_\rho^2\sqrt{2T_\circ(b+\Delta_{\max})\Lambda} + 4E_\rho\epsilon T_\circ + v_\rho T_\circ.\end{aligned} \tag{58}$$

The first line is the discretisation error. The second line controls Brownian fluctuations and the approximation by a finite net. The third line controls the centred time average, its net approximation and the stationary tail outside K_ρ.

Before giving the proof, let us describe its structure. The excess empirical risk $\widehat{\mathcal{R}}_n(\widetilde{G}) - \widehat{\mathcal{R}}_n(G)$ splits into three parts: a time average of $\|\widetilde{G} - G\|^2$ along the grid, a stochastic integral against the Brownian increments, and a discretisation error due to the grid. The discretisation error is controlled by the modulus ω_G (Steps 1–2). We then truncate to the ball K_ρ and replace the class by a finite net (Step 3). The time average concentrates around $\mathcal{R}_T$ thanks to the exponential β-mixing of Theorem 5, via the independent-block argument of Yu (1994) and Hoeffding's inequality (Step 4), while the Brownian term has Gaussian conditional increments and is controlled by a Chernoff bound (Step 5). A union bound over the net and the margin argument of Theorem 2 conclude (Step 6).

Proof of Theorem 4. Step 1: The drift modulus needs no growth assumption. Let G_A be the projection of G onto the closed Euclidean ball of radius A in $\mathbb{R}^D$. It is continuous and bounded, while $\|G - G_A\|_{L^2(\mu_\infty^q)} \to 0$ by Equation 48. Stationarity and the triangle inequality in L^2 give

$$\|G(Z_u) - G(Z_0)\|_{L^2(\mathbb{P}_\pi)} \le 2\|G - G_A\|_{L^2(\mu_\infty^q)} + \|G_A(Z_u) - G_A(Z_0)\|_{L^2(\mathbb{P}_\pi)}.$$

For fixed A, continuity of the paths and bounded convergence make the last term tend to zero as $u \downarrow 0$. First choosing A large and then u small proves $\omega_G(h) \to 0$ as $h \downarrow 0$; also $\omega_G(h) \le 2m_G$.

Fix a deterministic internal net $G_1, \dots, G_N$ on K_ρ and write $h = \widetilde{G} - G$ and $h_i = G_i - G$. The suprema below are measurable: the continuous drifts form a separable space in the topology of uniform convergence on compact sets, and their population risks are continuous on the class by domination with $\mathcal{E}^2$. One may therefore evaluate each supremum over a countable dense subset of the class.

Step 2: Exact expansion and discretisation. Set $\Delta W_k := W_{t_{k+1}} - W_{t_k}$. Inserting the SDDE increment into the square and cancelling the candidate-independent squared-increment term gives

$$\begin{aligned}\widehat{\mathcal{R}}_n(\widetilde{G}) - \widehat{\mathcal{R}}_n(G) - \mathcal{R}_T(\widetilde{G}) = A_n(h) - M_n(h) - D_n(h), \\ A_n(h) := \sum_k \Delta_k\left(\|h(Z_k)\|^2 - \int\|h(z)\|^2\,\mu_\infty^q(dz)\right), \\ M_n(h) := 2\sum_k\langle h(Z_k), \Sigma\Delta W_k\rangle, \\ D_n(h) := 2\sum_k\int_{t_k}^{t_{k+1}}\langle h(Z_k), G(Z_s) - G(Z_k)\rangle\,ds.\end{aligned} \tag{59}$$

Stationarity makes the centring in A_n exact; there is no further Riemann-sum error for its expectation. For every candidate,

$$|D_n(h)| \le \mathcal{D}_n := 2\sum_k\int_{t_k}^{t_{k+1}}\mathcal{E}(Z_k)\|G(Z_s) - G(Z_k)\|\,ds.$$

Cauchy–Schwarz and stationarity imply

$$\mathbb{E}_\pi\mathcal{D}_n \le 2m_{\mathcal{E}}\sum_k\int_0^{\Delta_k}\omega_G(u)\,du \le 2m_{\mathcal{E}}T_\circ\omega_G(\Delta_{\max}).$$

Markov's inequality therefore bounds $\sup_h |D_n(h)|$ by the first line of Equation 58, outside an event of probability at most η. If this expectation bound is zero, $\mathcal{D}_n = 0$ almost surely and no exceptional event is needed.

Step 3: Localisation and the squared-loss net. Let

$$\mathcal{L}_\rho := \left\{ \sup_{\tau_{\max} \le t \le T} \|Z_t - c^{(q)}\| \le \rho \right\}.$$

Since $\|Z_t - c^{(q)}\| \le \sqrt{q+1}\|X_t^{\mathrm{seg}} - c\|_{\mathcal{C}}$, a cover of $[\tau_{\max}, T]$ by J intervals of length at most one, stationarity and Markov's inequality give

$$\mathbb{P}_\pi(\mathcal{L}_\rho^c) \le \frac{J(q+1)^p S_p}{\rho^{2p}} \le \eta. \tag{60}$$

Define $f_h(z) := \|h(z)\|^2 \mathbf{1}_{K_\rho}(z) \in [0, E_\rho^2]$ and

$$A_n^\rho(h) := \sum_k \Delta_k \left(f_h(Z_k) - \int f_h(z)\, \mu_\infty^q(dz) \right).$$

On $\mathcal{L}_\rho$, the empirical squared losses equal their truncated versions, so $|A_n(h)| \le |A_n^\rho(h)| + T_\circ v_\rho$. For each h, choose a centre h_i with $\sup_{K_\rho} \|h - h_i\| \le \epsilon$. Both errors are bounded by E_ρ on this set, hence $\sup_z |f_h(z) - f_{h_i}(z)| \le 2E_\rho \epsilon$ and

$$\sup_h |A_n(h)| \le \max_{i \le N} |A_n^\rho(h_i)| + 4E_\rho \epsilon T_\circ + T_\circ v_\rho \quad \text{on } \mathcal{L}_\rho. \tag{61}$$

Step 4: Time-average fluctuations on an irregular grid. Partition the used left endpoints into bins $I_j = [\tau_{\max} + jb, \tau_{\max} + (j+1)b)$, for $0 \le j < J_b := \max\{1, \lceil T_\circ / b \rceil\}$, and set $w_j := \sum_{k : t_k \in I_j} \Delta_k$. The intervals with left endpoints in a nonempty bin are consecutive, and only the last can extend beyond its right boundary, by at most $\Delta_{\max}$. Consequently,

$$0 \le w_j \le b + \Delta_{\max}, \qquad \sum_j w_j = T_\circ, \qquad \sum_j w_j^2 \le (b + \Delta_{\max}) T_\circ. \tag{62}$$

Successive nonempty bins of the same parity have their observations separated by at least b in physical time. For either parity, Equation 50 bounds the total-variation distance between its joint block law and the product of its block marginals by $(m-1)_+ C_\beta e^{-\lambda_{\mathrm{erg}} b}$, where m is the number of its nonempty bins. Indeed, successively separating the next block from all previous blocks costs at most the corresponding β coefficient, and the triangle inequality adds these costs. This comparison concerns the observation blocks themselves, so its error is not multiplied by the number N of candidate functions. It is the independent-block argument of Yu (1994), here applied in physical time.

Under the product law for one parity, the variables

$$V_{j,i} := \sum_{k : t_k \in I_j} \Delta_k \left(f_{h_i}(Z_k) - \int f_{h_i}(z)\, \mu_\infty^q(dz) \right)$$

are independent and centred, with range length at most $E_\rho^2 w_j$. Hoeffding's inequality gives, for $x > 0$,

$$\mathbb{P}_{\mathrm{prod}} \left(\left| \sum_{j \text{ of one parity}} V_{j,i} \right| > x \right) \le 2 \exp\left(- \frac{2x^2}{E_\rho^4 \sum_j w_j^2} \right).$$

An empty parity has sum zero and needs no bound. Using Equation 62, taking $x = E_\rho^2 \sqrt{T_\circ (b + \Delta_{\max}) \Lambda / 2}$, and taking the union over the N centres and two parities yield

$$\mathbb{P}_\pi \left(\max_{i \le N} |A_n^\rho(h_i)| > E_\rho^2 \sqrt{2 T_\circ (b + \Delta_{\max}) \Lambda} \right) \le 4N e^{-\Lambda} + J_b C_\beta e^{-\lambda_{\mathrm{erg}} b} \le 2\eta. \tag{63}$$

The last inequality uses $b \ge 1$, hence $J_b \le J$, and Equation 57. Together with Equation 61, this supplies the third line of Equation 58.

Step 5: Brownian fluctuations and their net approximation. Use the predictable truncation

$$M_n^\rho(h) := 2\sum_k \langle h(Z_k)\mathbf{1}_{K_\rho}(Z_k), \Sigma\Delta W_k\rangle.$$

For fixed h_i, conditional on the past at t_k, the kth summand is centred Gaussian with variance at most $4\sigma_{\mathrm{op}}^2 E_\rho^2 \Delta_k$. Iterating the conditional moment-generating functions gives

$$\mathbb{E}_\pi e^{sM_n^\rho(h_i)} \le \exp\left(2s^2\sigma_{\mathrm{op}}^2 E_\rho^2 T_\circ\right), \qquad s \in \mathbb{R}.$$

The Gaussian Chernoff bound and a union over the fixed net imply

$$\mathbb{P}_\pi\left(\max_{i\le N}|M_n^\rho(h_i)| > 2\sigma_{\mathrm{op}}E_\rho\sqrt{2T_\circ\Lambda}\right) \le 2Ne^{-\Lambda} = \eta/2. \tag{64}$$

The net remainder is controlled pathwise, not by treating a data-selected remainder as a fixed martingale:

$$|M_n^\rho(h) - M_n^\rho(h_i)| \le 2\epsilon\sum_k \|\Sigma\Delta W_k\|, \qquad \mathbb{E}_\pi\sum_k\|\Sigma\Delta W_k\| \le \sigma_F\sum_k\sqrt{\Delta_k}.$$

Markov's inequality therefore bounds the remainder simultaneously for all candidates by $2\epsilon\sigma_F\eta^{-1}\sum_k\sqrt{\Delta_k}$, outside an event of probability at most η. On $\mathcal{L}_\rho$, $M_n(h) = M_n^\rho(h)$ for every h, so this and Equation 64 give the second line of Equation 58.

Step 6: Uniform deviation and mask recovery. The exceptional probabilities in Steps 2–5 sum to at most

$$\eta + \eta + 2\eta + \eta/2 + \eta = \tfrac{11}{2}\eta < \delta.$$

Adding the three bounds in Equation 59 proves Equation 51 on the complementary event. On that event, feasibility of (G, M) and empirical optimality imply

$$\mathcal{R}_T(\widehat{G}) + \lambda\sum_{a=0}^{q}\|\widehat{M}^a\|_0 - \lambda\sum_{a=0}^{q}\|M^a\|_0 \le \varepsilon_n(\delta;\rho,\epsilon).$$

The population margin of Theorem 2, applied to $\mu_\infty^q = \bar{\mu}_T^q$, makes the left-hand side at least $\gamma(\lambda)$ for every wrong mask. This uses its lower bound for every feasible wrong mask, not a population-minimisation assumption on the empirical estimator. Condition Equation 53 therefore excludes every wrong mask on the same event. Once the masks agree, their penalties cancel, leaving $\mathcal{R}_T(\widehat{G}) \le \varepsilon_n(\delta;\rho,\epsilon)$. Dividing by $T_\circ$ proves Equation 54. □

We showed in the proof that the discretisation term is $2m_{\mathcal{E}}T_\circ\omega_G(\Delta_{\max})/\eta$. When the stronger estimate $\omega_G(h) \le C_G\sqrt{h}$ is available, Step 2 improves this term to $4m_{\mathcal{E}}C_G(3\eta)^{-1}\sum_k\Delta_k^{3/2}$. It does not become $O(T_\circ\Delta_{\max})$ merely from a fourth drift moment. On an equally spaced used grid, we have the equality $\sum_k\sqrt{\Delta_k} = T_\circ/\sqrt{\Delta}$, so the martingale net remainder divided by $T_\circ$ is proportional to $\epsilon/(\eta\sqrt{\Delta})$. On a general grid, the displayed sum must be retained; it is not bounded using $\Delta_{\max}$ alone. The factors $1/\eta = 6/\delta$ together with the local envelope E_ρ, which may grow with the radius $\rho \ge (J(q+1)^p S_p/\eta)^{1/(2p)}$ and hence with $T_\circ$, make Equation 58 a qualitative consistency statement rather than a sharp rate.

Optimization accuracy and sampling endpoints. If the returned feasible pair has a certified empirical objective gap at most $\zeta_n \ge 0$, replace Equation 53 by $\varepsilon_n + \zeta_n < \gamma(\lambda)$ and the prediction bound by $(\varepsilon_n + \zeta_n)/T_\circ$. This follows by adding ζ_n to the final optimality comparison; an arbitrary output of a nonconvex optimization algorithm is not automatically certified. If $\tau_{\max}$ is not observed, let a be the first used left endpoint, replace $T_\circ$ by $T-a$, and use $\mathcal{R}_{a,T}(\widetilde{G}) := (T-a)\|\widetilde{G}-G\|^2_{L^2(\mu_\infty^q)}$ throughout this section and in its population margin. The proof is otherwise unchanged. Interpolated lag values require an additional error analysis and are not covered by the exact-observation statement above.

A.9 DISSIPATIVITY AND ERGODICITY OF THE LEARNED SYSTEM

In this section, we analyze the ergodicity of both the data-generating process and the learned simulator, providing theoretical insights beyond the empirical assumptions made in our experiments. Ergodicity of the true data-generating process guarantees that a single, long trajectory visits all states in proportion to their true probabilities. This allows us to learn the invariant measure because time averages converge to ensemble averages. Conversely, ergodicity of the learned model ensures stability during forward simulation: rather than drifting indefinitely or exploding, the generated paths settle into their own stationary distribution (which ideally matches the true one). In what follows, we discuss sufficient conditions and architectural choices to guarantee the ergodicity of the learned system.

We parameterize the drift using an MLP with $\tanh$ hidden activations and an affine output layer. Provided there are no skip connections to the final layer, the output of this architecture is globally bounded. Consequently, when used as the drift in Equation 1, the resulting process is non-explosive. However, a globally bounded drift cannot satisfy the dissipation condition of Assumption 6 outside a compact set, nor can it uniformly approximate a drift that does over the entire domain. As a result, Theorem 6 does not apply to an unconstrained setup. We emphasize that this highlights a limitation of the sufficient condition rather than the model itself, as bounded drifts can still be ergodic. For instance, the SDE $dX_t = -\tanh(X_t)\,dt + \sigma\,dW_t$ yields an integrable invariant density proportional to $(\cosh x)^{-2/\sigma^2}$, and its drift can be exactly represented by our unconstrained chosen architecture.

In what follows, we describe several fixes and results to maintain dissipativity and thus ergodicity of the learned drift. We chose for our experiments to apply the post-training exterior confinement defined by Equation 67 which imposes no restriction to the architecture during the training and only applies a restoring force outside a compact set that encompasses all of our samples.

Although an unconstrained MLP architecture does not inherently satisfy these global dissipation conditions, we nevertheless provide a conditional ergodicity result in Theorem 6. This global perturbation result becomes highly relevant when the learned and true drifts exhibit compatible tail behaviors — for instance, when they share a dissipative physical baseline and differ only by bounded residuals. Thus, we retain this theorem to address scenarios equipped with such structural priors, with the caveat that its hypotheses do not automatically emerge simply from fitting an unconstrained neural network.

Theorem 6 (Conditional ergodicity transfer under bounded drift error)**.** *Suppose Assumptions 1, 2 and 6 hold for the true system, with $B = \sum_{a=1}^{q}(\beta_a + \varepsilon_a) < \beta_0$. Let $\widetilde{G}$ be locally Lipschitz, of at most linear growth, and satisfy*

$$\sup_{z\in\mathbb{R}^{D(q+1)}} \|\widetilde{G}(z) - G(z)\| \le \delta_0 < \infty.$$

For any $0 < \eta < 2(\beta_0 - B)$, it satisfies Assumption 6 with the same centre and radius and with

$$\widetilde{\beta}_0 = \beta_0 - \eta/2, \qquad \widetilde{\alpha} = \alpha + \delta_0^2/(2\eta), \qquad \widetilde{\beta}_a = \beta_a,$$
$$\widetilde{C}_R = C_R + \delta_0^2/(2\eta) + \eta R^2/2, \qquad \widetilde{\varepsilon}_a = \varepsilon_a \quad (a \ge 1).$$

With the same invertible additive diffusion, the learned SDDE is non-explosive and its segment process has a unique invariant probability measure, all polynomial stationary segment moments, and geometric convergence in total variation as in Theorem 3, with model-dependent constants.

Proof. Writing $u = x_0 - c$, Cauchy–Schwarz and Young's inequality give

$$\langle u, \widetilde{G}(z) - G(z)\rangle \le \delta_0\|u\| \le \frac{\eta}{2}\|u\|^2 + \frac{\delta_0^2}{2\eta}.$$

Add this to the exterior and interior inequalities of Assumption 6; in the latter, use $\|u\|^2 \le R^2$. The displayed constants follow and retain strict domination. The resulting radial bound also gives the Khasminskii condition. Local Lipschitzness gives well-posedness by localisation, and the second-moment estimate together with at-most-linear growth gives finite drift energy for square-integrable initial histories independent of the future noise. The moment and kernel-overlap arguments establishing Theorem 3 then apply to $\widetilde{G}$. □

An important observation is that the constant δ_0 need not be small for existence of the ergodic behavior, although increasing it can have drawbacks on statistical accuracy, and worsen moments and mixing results. Note that the result above can be strengthened into an affine error bound with sufficiently small slope, $\|\widetilde{G}(z) - G(z)\| \le \delta_0 + \theta\|z - c^{(q)}\|$, $\theta \ge 0$.

For the learned process to remain stable, we only need to bound the network's error in the outward radial direction. Large errors that act tangentially, or those that provide additional inward pull, will not cause the system to diverge. We therefore turn to conditions that can be imposed directly on the learned drift rather than on its unknown error at infinity.

Dissipative architecture with bounded neural residual A simple and natural extension of the unconstrained architecture is to retain the MLP with tanh activations as nonlinear residuals and add an explicit instantaneous damping:

$$\widetilde{G}_\theta(z) = -K(x_0 - c) + N_\theta(z), \quad K = \mathrm{diag}(\kappa_1, \dots, \kappa_D),\ \kappa_i \ge \kappa_* > 0 \tag{65}$$

Proposition 2 (Confinement by construction)**.** *Let N_θ be locally Lipschitz and globally bounded, with $\sup_z \|N_\theta(z)\| \le M_\theta$. For any fixed delays and an invertible additive diffusion as in Assumption 1, the drift in Equation 65 defines a non-explosive SDDE whose segment process has a unique invariant law, all polynomial stationary segment moments, and geometric convergence in total variation.*

Proof. For every lagged state z, with $u = x_0 - c$,

$$\langle u, \widetilde{G}_\theta(z)\rangle \le -\kappa_*\|u\|^2 + M_\theta\|u\| \le -\frac{\kappa_*}{2}\|u\|^2 + \frac{M_\theta^2}{2\kappa_*}. \tag{66}$$

This gives Assumption 6 with zero delayed coefficients; the interior radial bound follows by increasing its constant on any chosen ball. Local Lipschitzness and linear growth ensure the remaining well-posedness and energy conditions. The conclusion is a direct application of Theorem 3. □

Post-training exterior confinement The same idea can be applied after training in such a way that the fitted drift remains unchanged inside a radius vector $R = (R_1, \dots, R_D)^\top$ with strictly positive components, $R_j > 0$. Set $u = x_0 - c$ and define $\Pi_R(u) = (\min\{R_j, \max\{-R_j, u_j\}\})_j$ the projection onto the box $\prod_j [-R_j, R_j]$. Then, define the post-hoc network:

$$\widetilde{G}_{\theta,R}(z) = N_\theta(z) - K \cdot (u - \Pi_R(u)) \tag{67}$$

Whenever $|x_{0,j} - c_j| \le R_j$ for every j, the network is left unchanged for all delayed inputs. This applies a linear restoring drift only to coordinates outside the given box. Furthermore, since we can rearrange Equation 67 as $\widetilde{G}_{\theta,R}(z) = -K\cdot u + (N_\theta(z) + K\cdot\Pi_R(u))$ and $\sup_z\|N_\theta(z) + K\Pi_R(u)\| \le M_\theta + \|K \cdot R\|$, Proposition 2 applies. Note, nonetheless, that this changes in general the invariant law of the learned process and thus the damping matrix and radii considered should be taken into consideration and their effect on long-run statistics assessed.

The preceding results describe some constructions to exhibit the existence of the learned equilibrium, however they say little about whether or not it agrees with the true one. Under the same invertible additive noise, we can find a stationary error bound on the invariant distributions in total variation given the true invariant measure.

Proposition 3 (Stationary error bound on invariant distributions)**.** *Suppose the true process is stationary with segment law π^{seg}. Let $\widetilde{G}$ be locally Lipschitz and define a non-explosive SDDE with the same lags and diffusion. Suppose its invariant law $\widetilde{\pi}^{\mathrm{seg}}$ and kernels obey*

$$\|\widetilde{P}_t(\varphi, \cdot) - \widetilde{\pi}^{\mathrm{seg}}\|_{\mathrm{TV}} \le \widetilde{C}e^{-\widetilde{\gamma}t}(1 + V(\varphi)), \quad V(\varphi) = \|\varphi - c\|_{\mathcal{C}}^2, \quad \pi^{\mathrm{seg}}V < \infty. \tag{68}$$

where the former inequality holds when the learned drift satisfies for example Assumptions 1, 2 and 6 as in Theorem 3. If $e_\Sigma^2 = \int\|\Sigma^{-1}(\widetilde{G} - G)(z)\|^2\mu_\infty^q(dz) < \infty$, then for every $t > 0$, with total variation normalized as $\sup_A|\nu(A) - \mu(A)|$,

$$\|\pi^{\mathrm{seg}} - \widetilde{\pi}^{\mathrm{seg}}\|_{\mathrm{TV}} \le \frac{e_\Sigma\sqrt{t}}{2} + \widetilde{C}e^{-\widetilde{\gamma}t}(1 + \pi^{\mathrm{seg}}V). \tag{69}$$

Proof. Step 1: Common starting equilibrium law. Fix $t > 0$ and start both equations with an initial segment of law π, independent of future noise. The true terminal-segment law remains π, whereas the learned law is $\pi\widetilde{P}_t$. Hence

$$\|\pi^{\mathrm{seg}} - \widetilde{\pi}^{\mathrm{seg}}\|_{\mathrm{TV}} \le \underbrace{\|\pi^{\mathrm{seg}} - \pi^{\mathrm{seg}}\widetilde{P}_t\|_{\mathrm{TV}}}_{\text{finite-time drift comparison}} + \underbrace{\|\pi^{\mathrm{seg}}\widetilde{P}_t - \widetilde{\pi}^{\mathrm{seg}}\|_{\mathrm{TV}}}_{\text{learned mixing}}. \tag{70}$$

Let $\mathbb{P}$ and $\mathbb{Q}$ be the true and learned laws of the full path on $[-\tau_{\max}, t]$, including its initial history. On this path space set

$$v(z) = \Sigma^{-1}(G(z) - \widetilde{G}(z)), \qquad A_s = \int_0^s \|v(Z_r)\|^2\, dr.$$

Stationarity under $\mathbb{P}$ gives the essential identity

$$\mathbb{E}_{\mathbb{P}} A_t = \int_0^t \mathbb{E}_{\mathbb{P}}\|v(Z_s)\|^2\, ds = te_\Sigma^2. \tag{71}$$

Step 2: Stopped change of drift. Finite mean energy does not imply a global Novikov condition. Instead, stop at $\sigma_m = \inf\{s \in [0,t] : A_s \ge m\}$, with the convention $\inf \varnothing = \infty$. Under $\mathbb{Q}$, let $\widetilde{B}$ be the driving Brownian motion and define

$$\frac{d\mathbb{P}_m}{d\mathbb{Q}} = \exp\left\{\int_0^{t\wedge\sigma_m} v(Z_s)\cdot d\widetilde{B}_s - \frac{1}{2}A_{t\wedge\sigma_m}\right\}. \tag{72}$$

Because $A_{t\wedge\sigma_m} \le m$, this exponential is a true martingale, also conditionally on the initial history. Girsanov's theorem therefore preserves the initial law π and makes $\mathbb{P}_m$ follow drift G until σ_m, then drift $\widetilde{G}$ afterwards.

Under $\mathbb{P}_m$, the log-density in Equation 72 is a mean-zero stochastic integral plus $A_{t\wedge\sigma_m}/2$. Consequently, the Kullback-Leibler divergence can be written as

$$D_{\mathrm{KL}}(\mathbb{P}_m\|\mathbb{Q}) = \frac{1}{2}\mathbb{E}_{\mathbb{P}_m} A_{t\wedge\sigma_m} = \frac{1}{2}\mathbb{E}_{\mathbb{P}} A_{t\wedge\sigma_m} \le \frac{te_\Sigma^2}{2}. \tag{73}$$

The second equality holds because the auxiliary and true equations agree up to the stopping time: pathwise uniqueness from Proposition 1 gives identical stopped path laws. This is why the error remains evaluated under the *true* law.

If we couple these two equations with the same history and Brownian motion, their paths can differ only if $\sigma_m \le t$, so

$$\|\mathbb{P} - \mathbb{P}_m\|_{\mathrm{TV}} \le \mathbb{P}(A_t \ge m) \le \frac{te_\Sigma^2}{m}.$$

Pinsker's inequality and Equation 73 now imply

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} \le \frac{te_\Sigma^2}{m} + \sqrt{\frac{1}{2}D_{\mathrm{KL}}(\mathbb{P}_m\|\mathbb{Q})} \le \frac{te_\Sigma^2}{m} + \frac{e_\Sigma\sqrt{t}}{2}.$$

Let $m \to \infty$ with t fixed. Extracting a terminal segment cannot increase total variation, and its laws under $\mathbb{P}, \mathbb{Q}$ are $\pi^{\mathrm{seg}}, \pi^{\mathrm{seg}}\widetilde{P}_t$, respectively. Thus

$$\|\pi^{\mathrm{seg}} - \pi^{\mathrm{seg}}\widetilde{P}_t\|_{\mathrm{TV}} \le \frac{e_\Sigma\sqrt{t}}{2}. \tag{74}$$

Step 3: Conclusion. Integrating Equation 68 against its initial law π^{seg} gives

$$\|\pi^{\mathrm{seg}}\widetilde{P}_t - \widetilde{\pi}^{\mathrm{seg}}\|_{\mathrm{TV}} \le \int \|\widetilde{P}_t(\varphi,\cdot) - \widetilde{\pi}^{\mathrm{seg}}\|_{\mathrm{TV}}\, \pi^{\mathrm{seg}}(d\varphi) \le \widetilde{C}e^{-\widetilde{\gamma}_t}(1 + \pi^{\mathrm{seg}}V).$$

Together with Equation 74 and Equation 70, this proves Equation 69. □

B COMPARISON WITH ADAPTIVE GROUP LASSO ASSUMPTIONS

Bellot et al. (2022) prove local consistency of the *adaptive group lasso*, for Markovian systems. They show similar identifiability proofs using the adaptive group lasso to select input features. AGL also requires an initial consistent estimator to build the adaptive weights. In this work, we derive identifiability guarantees for SDDE-driven systems, which extend Bellot et al. (2022) results. Additionally, we use the L_0 penalty instead of AGL. The L_0 route is in two ways cleaner: no initial estimator is needed, and the population theorem (2) already supplies an explicit λ window. Both results suppose an optimization oracle, which is standard in the literature, but something that is most often not available in practice.

Our theorems assume global minimization of a nonconvex, combinatorial objective. Bellot et al. (2022) assume that they can find exact solutions to their problem, which is least squares over a *neural network* class with an AGL penalty, and hence also nonconvex. Both results presuppose an optimization oracle — theirs for a continuous nonconvex problem, ours for a combinatorial one —. Empirically, we find that even on Markovian systems (part of the CausalDynamics benchmark, the Dysts benchmark), the relaxed L_0penalty performs better than the AGL-penalized objective. Our minimum-signal condition $s_{\min} > 0$ plays exactly the role of their eq. (10), and our effective sample size $\lambda_{\text{erg}}T$ plays the role of their $n/\|\alpha\|_2$.

C IMPLEMENTATION DETAILS AND HYPERPARAMETERS

C.1 IMPLEMENTATION DETAILS

The L_0-norm is not differentiable. To allow for score-based optimization, we implement the relaxed L_0 penalization proposed by Louizos et al. (2018). More precisely, the matrix M is a square logits matrix parametrized as the sigmoid of differentiable parameters in $\mathbb{R}$. During training, at each batch, we draw samples of M through a hard-concrete distribution, a continuous relaxation of a binary random variable. We use a low temperature parameter $\beta = 0.33$ to sharpen the hard-concrete distribution and force to turn inputs on or off. Samples are drawn on the $[-0.1, 1.1]$ interval before being clamped to $[0, 1]$ i.e. most samples will be exactly equal to 0 or 1. At inference, we fix M to be 0 or 1 if the probabilities (obtained from the logits) are lower or higher than 0.5. To have a smoothed convergence, we draw multiple samples of M at each batch before optimizing for its differentiable parameters. We also warm-up the model and train the neural network without penalty for 100 warm-up iterations before adding the penalty to the loss. We then train until the loss plateaus and does not decrease for 200 patience iterations. These parameters and optimization procedure are standard when using L_0-penalty (Louizos et al., 2018).

The functions f are small multi-layer perceptrons (MLPs) and we use a gradient penalty to enforce that our functions are Lipschitz. To make sure that smooth dimensions e.g. with close to 0 gradient are not approximated by the mean and then do not have any drivers, we renormalize each dimension by the gradient mean and standard deviation.

C-NODE follows a similar implementation, where the L_0 is replaced by L_1norm. In practice, we just apply a L_1penalty on the logits matrix M. AGL implementation is similar, but uses an AGL penalty instead. We find that using a gradient penalty slightly improve the results for both, so we included it in the method's implementation. Except the penalty, everything is kept the same between methods.

D CAUSALDYNAMICS BENCHMARK

D.1 METRIC COMPUTATION

L_0-NDDE and "No Grad. Pen." use a logit matrix to represent probabilities of inputting a feature or not. It is thus straightforward to calculate SHD, AUROC and AUPRC by comparing this matrix to the ground truth logits matrix. AGL and L_1penalty instead assign weights to each input feature. While these weights can still be compensated for by the neural network, we used this weight matrix to compute AUROC / AUPRC, treating the weights as feature importance. For SHD instead, we used a low threshold (10^{-8}) to decide whether an input feature is turned off or not. Changing this

threshold did not affect performance, as we observed that the weights were generally either 0 or above 10^{-3} in practice.

D.2 HYPERPARAMETER SEARCH

The causal methods evaluated in CausalDynamics have various numbers of hyperparameters. While methods such as PCMCI rely crucially on a single parameter e.g. the conditional independence test threshold, others such as TSCI or NGC depend on many parameters e.g. neural network architectures. Herdeanu et al. (2025) thus decided to keep default parameters except the main parameter and perform hyperparameter search on this parameter only. We do the same and keep default parameters, highlighted in Section D.3, except for the sparsity penalty coefficient, which controls for the final level of sparsity.

For all methods, we run our model on the 10 time series of the Lorenz84 system with the following penalty coefficients and select the one maximizing AUROC: (0.0001, 0.00025, 0.0005, 0.001, 0.0025, 0.005, 0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1, 2.5, 5, 10, 25, 50, 100).

D.3 FINAL HYPERPARAMETERS

Parameter	Value
Architecture, functions f_j	
Number of layers	2
Number of units per layer	8
Optimization parameters	
Gradient penalty coefficient	100
Warm-up iterations	100
Patience iterations	200
Learning rate	0.001
Batch size	128
L_0 implementation parameters (L_0-NDDE, No grad. pen.)	
Temperature β_{temp}	0.33
Approximate distribution interval	$[-0.1, 1.1]$
Number of samples	3
Sparsity penalty coefficients	
L_0	0.025
AGL	5
C-NODE (L1)	50
No grad. pen.	0.01

Table 3: **Final hyperparameters used for L_0-NDDE, C-NODE, AGL and No grad. pen.** The last 4 lines show the penalty coefficient for each method, as it is the only parameter tuned for each the method. Note that No grad. pen. does not use gradient penalty.

D.4 FULL CAUSALDYNAMICS RESULTS

	Experiment	Simple			Coupled					Climate		Avg. rank
		Default	Confounder	Noise	Default	Noise	Confounder	Time-lag	Standardize	MAOOAM	ENSO	
SHD	PCMCI+	41.04	23.02	47.64	224.80	183.90	324.63	327.72	228.32	80.00	529.36	7.30
	F-PCMCI	35.30	21.07	45.09	192.90	149.60	195.74	350.61	201.79	130.00	530.27	5.80
	VARLiNGAM	35.69	22.04	42.84	311.45	248.90	159.63	449.33	349.63	130.00	453.00	8.00
	DYNOTEARS	52.37	21.74	51.68	181.50	180.95	248.32	261.44	243.84	94.00	589.36	6.60
	NGC	28.91	19.96	28.84	840.95	842.55	670.53	793.67	840.26	31.00	337.09	7.90
	TSCI	52.50	22.02	60.50	174.70	173.50	265.26	244.56	244.42	108.00	666.27	7.20
	CUTS+	48.11	24.04	61.32	**152.00**	150.50	272.68	247.22	310.63	130.00	608.73	7.50
	RCD	61.85	26.74	61.38	157.05	155.65	136.53	**201.11**	**159.84**	130.00	665.36	6.80
	GRaSP	59.04	27.09	60.18	215.94	842.55	**136.00**	793.67	840.26	126.00	666.27	9.70
	TCDF	59.46	26.61	61.43	252.00	842.55	670.53	793.67	840.26	130.00	666.27	11.65
	AGL	26.60	16.96	21.28	319.04*	222.41*	246.37*	315.54*	312.31*	**30.00***	337.09	5.55
	C-NODE (L1)	**24.07**	17.10	**20.5**	253.91*	176.30*	179.58*	273.57*	258.09	42.00	335.36	4.50
	L_0-**NDDE (Ours)**	26.42	**16.25**	23.67	205.35	**144.20**	143.33	222.34	212.11	40.00	**325.54**	**2.50**
AUROC / AUPRC	PCMCI+	.52 / .71	.49 / .59	.50 / .69	**.67** / .25	.64 / .25	.58 / .20	.58 / .24	**.69** / .27	.69 / .88	.57 / .70	4.10 / 5.25
	F-PCMCI	.51 / .70	.50 / .59	.52 / .70	**.67** / .27	.57 / .21	.55 / .19	.59 / .24	.68 / .28	.50 / .81	.57 / .70	4.80 / 6.05
	VARLiNGAM	.50 / .69	.48 / .58	.53 / .70	.60 / .19	.57 / .18	.51 / .17	.54 / .22	.60 / .19	.50 / .81	.56 / .69	6.90 / 8.55
	DYNOTEARS	.43 / .67	.52 / .64	.48 / .68	.59 / .21	.57 / .20	.49 / .17	.53 / .22	.65 / .23	.64 / .86	.55 / .69	7.75 / 7.95
	NGC	.50 / .69	.50 / .58	.50 / .68	.50 / .15	.50 / .15	.50 / .16	.50 / .20	.50 / .15	.50 / .81	.50 / .67	9.85 / 10.90
	TSCI	.46 / .68	.53 / .66	.49 / .68	.60 / .23	.53 / .17	.51 / .18	.53 / .21	.65 / .23	.58 / .84	.50 / .67	7.50 / 8.00
	CUTS+	.50 / .69	.50 / .58	.50 / .68	.50 / .15	.50 / .15	.49 / .16	.50 / .20	.50 / .15	.50 / .81	.50 / .67	10.10 / 10.90
	RCD	.50 / .69	.50 / .58	.50 / .68	.50 / .15	.50 / .15	.51 / .18	.50 / .20	.50 / .16	.50 / .81	.50 / .67	9.55 / 10.25
	GRaSP	.52 / .71	.55 / .64	.49 / .68	.49 / .15	.50 / **.50**	.50 / .17	.50 / **.50**	.50 / **.50**	.48 / .81	.50 / .50	9.60 / 6.70
	TCDF	.51 / .70	.50 / .58	.50 / .68	.48 / .15	.50 / **.50**	.50 / **.50**	.50 / **.50**	.50 / **.50**	.50 / .81	.50 / .50	9.85 / 6.65
	AGL	.52 / .71*	.56 / .64	.53 / .74*	.53 / .33*	.53 / .33*	.54 / .35*	.52 / .35*	.54 / .34*	.56 / .86*	.54 / .69*	5.80 / 4.35
	C-NODE (L1)	.64 / .79	.54 / .65	.61 / .79*	.64 / **.43**	.63 / .43	**.67** / .47	.59 / .43	.68 / .47	**.88 / .97**	.49 / .66*	3.45 / 3.35
	L_0-**NDDE (Ours)**	**.64 / .80**	**.58 / .66**	**.63 / .80**	.64 / **.43**	**.65** / .44	**.67** / .47	**.60** / .43	.66 / .45	.86 / .96	**.58 / .75**	**1.75 / 2.10**

Table 4: **L_0-NDDE outperforms competing methods most of the time on the CausalDynamics benchmark.** This table is similar to Table 1 with additional methods being shown.

D.5 No gradient penalty

	Experiment	Simple			Coupled					Climate	
		Default	Confounder	Noise	Default	Noise	Confounder	Time-lag	Standardize	MAOOAM	ENSO
SHD	No Grad. Pen.	26.95	18.21	24.47	215.08	142.67	145.08	233.58	220.68	49.00	341.27
	L_0-NDDE (Ours)	26.42	16.25	23.67	205.35	144.20	143.33	222.34	212.11	40.00	325.54
AUROC / AUPRC	No Grad. Pen.	.64 / .80	.58 / .65	.61 / .79	.63 / .42	.64 / .44	.67 / .47	.59 / .41	.65 / .44	.85 / .96	.52 / .68
	L_0-NDDE (Ours)	.64 / .80	.58 / .66	.63 / .80	.64 / .43	.65 / .44	.67 / .47	.60 / .43	.66 / .45	.86 / .96	.58 / .75

Table 5: L_0-NDDE is slightly stronger with gradient penalty. Gradient penalty helps avoid overfitting spurious correlations and leads to slightly better performance on the CausalDynamics benchmark.

E Dysts benchmark

E.1 Train and test systems

We use 40 systems in total, split into a hyperparameter search set (5 systems), and a test set (the other 35 systems). For each system, we generate one training trajectory, and 3 test trajectories. The train trajectories last 50 Lyapunov times and the test trajectories last 250 Lyapunov times, to ensure that the trajectory visits the entire space. Trajectories are sampled at the native temporal resolution. To avoid too short or too long trajectories in training, we enforce the training trajectory to have a minimum of 1000 and a maximum of 10000 timesteps. We use a stochastic forcing of level 0.05 times the variance of each dimension of the training trajectory. Each test and training trajectory is normalized by subtracting the mean and dividing by the training trajectory standard deviation per dimension.

We studied 40 systems in total, which are the 40 Dysts systems for which the Jacobian is accessible. For these 40 systems, a method is implemented to access the Jacobian at any point along any trajectory. This is necessary to compute metrics. Among these 40 systems, we randomly chose 5 systems for doing hyperparameter search, and 35 systems for testing. On the 5 hyperparameter search systems, we ran the methods with the following sets of hyperparameters, and selected the ones that maximized the valid prediction time on the three test trajectories.

For all methods, we wait a certain number of patience iterations (200) to declare convergence i.e. when the penalized loss does not improve for 200 iterations, we stop sparsifying. After this, we fix the input mask i.e. the learned matrices $\tilde{M}$, and then keep training the neural networks to learn the dynamics for a certain number of iterations `n_iter_after_convergence`. Also, we wait a number `niter_warmup` of warm-up iterations before sparsifying.

We chose the best values among the following values for the following parameters:

- Sparsity penalty coefficient: $[0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100]$
- Gradient penalty coefficient: $[0, 10, 100, 1000, 10_000]$
- `n_iter_after_convergence`: $[0, 50, 100, 200, 500, 1000]$
- `niter_warmup`: $[0, 25, 50, 100]$

For all methods, these four parameters are key as they control the sparsity and the regularization of the methods and their final values can be found in Table 6. All other parameters are kept the same as in the CausalDynamics experiments (Table 6).

Optimization parameters	L_0-NDDE	NODE	C-NODE	AGL	PathReg
Gradient penalty coefficient	0	100	0	1000	1000
`niter_warmup`	0	0	100	50	200
`n_iter_after_convergence`	200	0	500	1000	500
Sparsity penalty coefficients	0.05	0	0.05	50	1

Table 6: **Final hyperparameters used for L_0-NDDE, C-NODE, AGL and PathReg**

The systems chosen randomly to perform hyperparameter search are the following: MooreSpiegel, Lorenz, Laser, RikitakeDynamo, Chua

The systems used for evaluating the methods are the following: BurkeShaw, Chen, ChenLee, Coullet, DequanLi, Duffing, Finance, GenesioTesi, Hadley, JerkCircuit, KawczynskiStrizhak, LiuChen, Lorenz84, LuChen, LuChenCheng, PanXuZhou, PehlivanWei, QiChen, RayleighBenard, Rucklidge, Sakarya, SanUmSrisuchinwong, ShimizuMorioka, SprottA, SprottB, SprottC, SprottE, SprottI, SprottTorus, Thomas, ThomasLabyrinth, Tsucs2, WangSun, YuWang, YuWang2, ZhouChen

E.2 METRICS

As we evaluate driver identifiability on the CausalDynamics benchmark, we are now interested in evaluating the dynamics. More precisely, we want to compare methods on short-term horizon prediction performance, attractor and geometry reconstruction, and long-term statistics. At evaluation, we predict autoregressively 3 trajectories from the initial conditions of the 3 test conditions (only the initial conditions are given), and generate trajectories of the same length as the test trajectories. We then compare the predicted vs true trajectories and compute several metrics.

For short-term prediction, we compute the valid prediction time at 1, normalized by the first Lyapunov exponent (valid prediction lyapunov time, VLPT) i.e. we compute the number of Lyapunov timesteps before the normalized trajectory is further than 1 away from the true normalized trajectory. We also compute the normalized root mean square error (NRMSE), by averaging the RMSE across the first 5 Lyapunov timesteps before normalizing by the trajectory mean and standard deviation.

For geometry and attractor reconstruction, we compute the first Lyapunov exponent error (FLEE) i.e. the squared difference between the first Lyapunov exponent of the learned vs. true system. The Lyapunov exponents are computed using the Benettin algorithm (Skokos, 2010), which requires access to the Jacobian which is possible in the selected Dysts systems, and for the learned models which are by construction differentiable. This error is small when the chaoticity of the system is well approximated. We additionally compute the Kaplan-Yorke dimension error (D_{KY}, Evans et al., 2000). Computed using the Lyapunov exponents, the D_{KY} gives us the dimension of the attractor i.e. the fractal dimension, which would be 2 in the case of the classic Lorenz system. The reported error is the squared error between the estimated system vs. true system's D_{KY}.

We then evaluate whether the learned system correctly captures long-term statistics, in other words whether the predicted trajectories converge, and if so whether they converge to the right equilibrium distribution. We report the log-spectral distance (LSD), which is the squared difference between the true and estimated log power spectral density, to indicate whether the system captured the right oscillation and periodicity properties of the system. We additionally report the D_{stsp}, a state-space

divergence which estimates the KL divergence between the learned and true system invariant distribution from a set of predicted and true trajectories (Koppe et al., 2019). We compute it by placing a Gaussian kernel at each trajectory point (for both true and generated trajectories) and estimating the KL divergence between the two Gaussian Mixture Models via Monte Carlo sampling.

E.3 RESULTS

On top of the quantitative results shown in the paper, we show, in Figure 2 the predicted (red) vs true (black) trajectories for 26 systems, for the three best methods, L_0-NDDE, NODE and C-NODE.

We then analyze the performance of the different methods on this benchmark, in Figure 3.

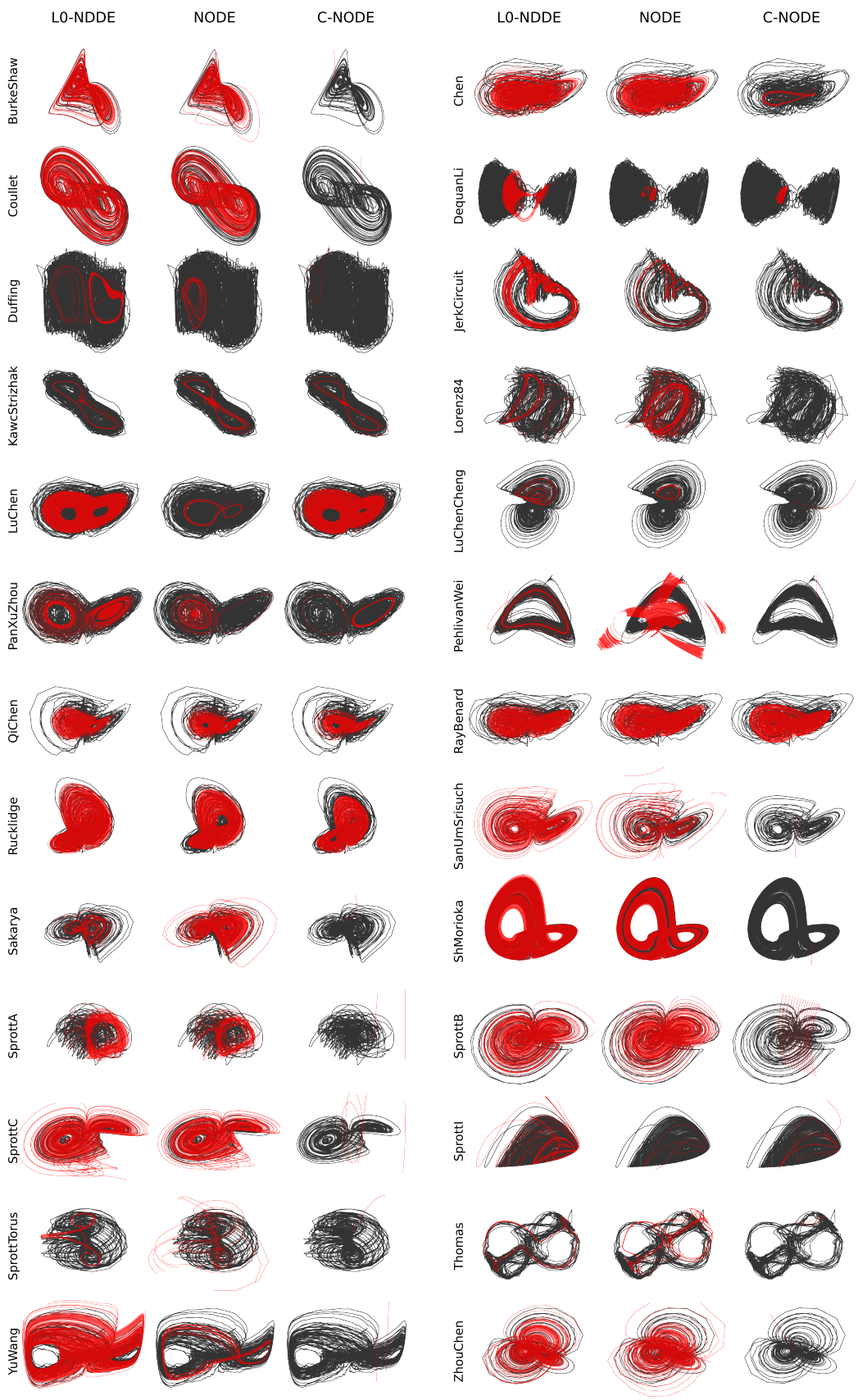

Figure 2: Example predicted trajectories (red) vs. true trajectories (black) on 26 Dysts systems, for L_0-NDDE (left), NODE (middle), and C-NODE (right).

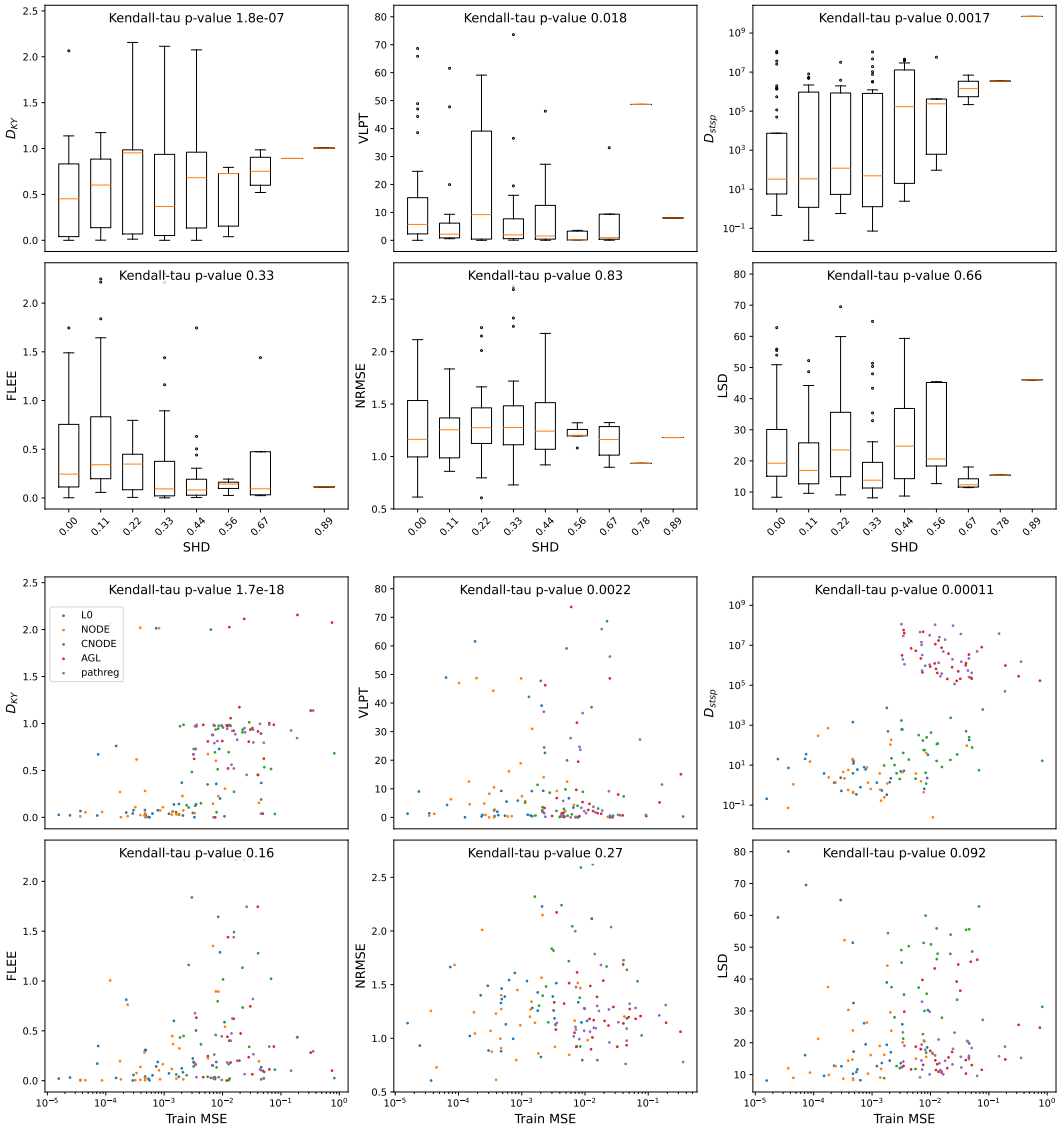

Figure 3: **Lower training error and lower SHD are both associated with higher performance.** We perform an analysis to better understand why L_0-NDDE and NODE outperform other methods on the Dysts datasets. We compute, for all methods and all datasets, the training error (Train MSE) and SHD. We then plot the 6 metrics versus the SHD (upper half) and versus the training error (lower half). We see that D_{KY} and D_{stsp} increase, while VLPT decreases with SHD and training error. These relationships are statistically significant according to a Kendall-τ test, a standard rank-correlation test for investigating associations between two sets of values, without assuming Gaussianity of the samples or needing large samples. The relationship is also significant between LSD and training error. This indicates that a lower training error is associated with better reconstruction, indicating why NODE performs well, as the absence of input feature regularization allows it to overfit to the training data; and why L_0-NDDE outperforms other methods, as it achieves lower SHD. Moreover, the L_0-penalty has the advantage that it does not perform weight decay, but explicitly turns features on or off, allowing the rest of the network to approximate the dynamics very well given the set of selected features.