Causal Modeling with Stationary Diffusions

Lars Lorch
Dept. of Computer Science
ETH Zürich, Switzerland
llorch@ethz.ch

Andreas Krause*
Dept. of Computer Science
ETH Zürich, Switzerland
krausea@ethz.ch

Bernhard Schölkopf*
MPI for Intelligent Systems
Tübingen, Germany
bs@tuebingen.mpg.de

Abstract

We develop a novel approach towards causal inference. Rather than structural equations over a causal graph, we learn stochastic differential equations (SDEs) whose stationary densities model a system's behavior under interventions. These stationary diffusion models do not require the formalism of causal graphs, let alone the common assumption of acyclicity. We show that in several cases, they generalize to unseen interventions on their variables, often better than classical approaches. Our inference method is based on a new theoretical result that expresses a stationarity condition on the diffusion's generator in a reproducing kernel Hilbert space. The resulting *kernel deviation from stationarity (KDS)* is an objective function of independent interest.

1 Introduction

Decision-making, e.g., in the life sciences, requires predicting the outcomes of *interventions* in a system $\mathbf{x} \in \mathbb{R}^d$. To achieve this, causal inference models $\mathbf{x}$ with a structural causal model (SCM) (Pearl, 2009)

$$\mathbf{x} = f(\mathbf{x}, \boldsymbol{\epsilon}) \,, \tag{1}$$

where $\epsilon_j \in \mathbb{R}$ are random noise variables, and often $x_j = f_j(\mathbf{x}) + \epsilon_j$. Interventions can be realized as modifications of the functions f_j or ϵ_j, and the SCM enables us to estimate the induced distribution shifts in $\mathbf{x}$. However, as x_j depends recursively on $\mathbf{x}$, SCMs are generally limited to modeling *acyclic* causal effects.

In this work, we propose to model a system's causal dependencies and their entailed probability distributions with stochastic differential equations (SDEs) and their entailed *stationary* densities. Specifically, we replace the SCM by its continuous-time stochastic analogue

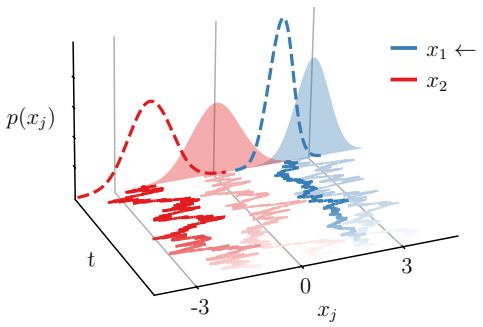


Figure 1: Stationary SDEs as causal models. Bottom axes show sample paths of a stationary diffusion in $\mathbb{R}^2$ before (pale) and after (dark) an intervention on x_1. The marginals $p(x_j)$ visualize the distribution shift.

$$\mathrm{d}\mathbf{x}_t = f(\mathbf{x}_t)\mathrm{d}t + \sigma(\mathbf{x}_t)\mathrm{d}\mathbb{W}_t \,. \tag{2}$$

Akin to real-world processes, SDEs unroll causal dependencies over time t, yet the densities modeled by stationary SDEs remain time-invariant, like the observations $\mathbf{x}$. Just as in SCMs, interventions may be modeled as modifications to f and σ; the SDEs then characterize how the stationary density of $\mathbf{x}$ changes by propagating the perturbations through its causal mechanisms (Figure 1).

In the following, we will argue that modeling causation using stationary diffusions has several benefits. Because dependencies get unrolled over time, SDEs can model feedback cycles, which SCMs allow only under strong model restrictions (e.g., Mooij et al., 2011; Rothenhäusler et al.,

Accepted to the NeurIPS 2023 Workshop on Causal Representation Learning.

*Equal supervision.

2015). Since acyclicity is not a constraint, our approach does not require constrained optimization or causal graphs. Moreover, the inference method we derive is agnostic to the system and intervention model, contrary to many SCM approaches (e.g., Shimizu et al., 2006). Our novel objective enables us to learn stationary SDEs via gradient-based optimization, without sampling from the model or backpropagating gradients through time.

2 Background

To describe our approach, we first review background on kernels, SDEs, and their generators.

Kernels and reproducing kernel Hilbert spaces Let $k(\mathbf{x}, \mathbf{x}') : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ denote a positive definite kernel function that is four times differentiable. Additionally, let $\mathcal{H}$ be the reproducing kernel Hilbert space (RKHS) of functions $\mathbb{R}^d \to \mathbb{R}$ associated with the kernel k and equipped with the norm $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The RKHS $\mathcal{H}$ satisfies that $k(\cdot, \mathbf{x}) \in \mathcal{H}$ for all $\mathbf{x} \in \mathbb{R}^d$, where $k(\cdot, \mathbf{x})$ denotes the function obtained when fixing the second argument of k at $\mathbf{x}$. The RKHS $\mathcal{H}$ also satisfies the *reproducing property* that $h(\mathbf{x}) = \langle h, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}}$ for all $\mathbf{x} \in \mathbb{R}^d$ and $h \in \mathcal{H}$. Thus, evaluations of RKHS functions $h \in \mathcal{H}$ are inner products in $\mathcal{H}$ and parameterized by $k(\cdot, \mathbf{x})$. (Schölkopf and Smola, 2002).

Stochastic differential equations SDEs are a stochastic analogue to differential equations. Rather than functions, their solutions are stochastic processes $\{\mathbf{x}_t\}$, $\mathbf{x}_t \in \mathbb{R}^d$ called diffusions. The Wiener process $\{\mathbb{W}_t\}$, $\mathbb{W}_t \in \mathbb{R}^b$ can be viewed as driving noise with independent increments $\mathbb{W}_{t+s} - \mathbb{W}_t \sim \mathcal{N}(0, s\mathbf{I})$, where usually $b = d$. General SDEs as in (2) contain a drift $f : \mathbb{R}^d \to \mathbb{R}^d$ and a diffusion function $\sigma : \mathbb{R}^d \to \mathbb{R}^{d \times b}$ with some $\mathbf{x}_0 \sim p_0$. We assume that f and σ are Lipschitz continuous, which ensures that the SDEs in (2) have a unique strong solution given the initial vector $\mathbf{x}_0$ (Øksendal, 2003, Theorem 5.2.1). The diffusion $\{\mathbf{x}_t\}$ solving the SDEs is *stationary* if the probability density $\mu_t(\mathbf{x})$ of $\mathbf{x}_t$ at time t is the same for all $t \geq 0$ (Ethier and Kurtz, 1986, Chapter 4, Lemma 9.1).

The infinitesimal generator The local evolution of a diffusion is described by its infinitesimal generator. The generator $\mathcal{A}$ associated to a stochastic process $\{\mathbf{x}_t\}$ is a linear *operator* that maps functions $h : \mathbb{R}^d \to \mathbb{R}$ to functions of the same signature. $\mathcal{A}$ can be viewed as the derivative of the semigroup of transition operators $\{\mathcal{T}_t : t \geq 0\}$ given by $(\mathcal{T}_t h)(\cdot) = \mathbb{E}_{\{\mathbf{x}_t\}}[h(\mathbf{x}_t) \,|\, \mathbf{x}_0 = \cdot]$:

$$(\mathcal{A}h)(\mathbf{x}) := \lim_{t \downarrow 0} \frac{\mathcal{T}_t h(\mathbf{x}) - h(\mathbf{x})}{t} \tag{3}$$

for all functions $h \in \mathrm{dom}(\mathcal{A})$. The *domain* $\mathrm{dom}(\mathcal{A})$ of the generator contains all functions for which this limit exists for all $\mathbf{x} \in \mathbb{R}^d$ (Ethier and Kurtz, 1986, Chapter 1.1). Intuitively, the generator tells us how $h(\mathbf{x}_t)$ changes infinitesimally over time t—in expectation and given an arbitrary function h. If the stochastic process $\{\mathbf{x}_t\}$ solves the SDEs (2), then its generator $\mathcal{A}$ can be expressed in terms of f and σ in the SDEs for a large class of functions h. Specifically, for all $h \in C_c^2$, we have $\mathcal{A} = \mathcal{L}$ and $h \in \mathrm{dom}(\mathcal{A})$, where $\mathcal{L}$ is the linear differential operator $\mathcal{L}$ given by (Øksendal, 2003, Theorem 7.3.3)

$$(\mathcal{L}h)(\mathbf{x}) := \; f(\mathbf{x}) \cdot \nabla_{\mathbf{x}} h(\mathbf{x}) + \tfrac{1}{2} \operatorname{tr}\big(\sigma(\mathbf{x})\sigma(\mathbf{x})^\top \nabla_{\mathbf{x}} \nabla_{\mathbf{x}} h(\mathbf{x})\big) \tag{4}$$

3 The Kernel Deviation from Stationarity

Given a target density μ, how can we learn the functions f and σ of a general system of SDEs (2) such that the diffusion solving the SDEs has the stationary density μ? In this first part, we will study this general inference question without yet considering causality and interventions in SDEs. Our starting point is a well-known link between the generator of a stochastic process and its stationary density. For a stochastic process $\{\mathbf{x}_t\}$, the density μ is the stationary density if and only if

$$\mathbb{E}_{\mathbf{x} \sim \mu}\big[\mathcal{A}h(\mathbf{x})\big] = 0 \tag{5}$$

for all functions h in a *core* for the generator $\mathcal{A}$ (Ethier and Kurtz, 1986, Chapter 4, Proposition 9.2). Roughly speaking, a core is a dense subset of functions in the domain $\mathrm{dom}(\mathcal{A})$ such that, if (5) holds for the core, then (5) also holds for all $h \in \mathrm{dom}(\mathcal{A})$ (see Hansen and Scheinkman, 1995). Equation (5) states that every function h of $\{\mathbf{x}_t\}$ must have zero rate of change $\mathcal{A}h(\mathbf{x})$, that is, must be invariant with time t, in expectation over the stationary density $\mathbf{x} \sim \mu$.

If we can verify that the expected infinitesimal change over a target density μ is zero for an expressive class of test functions h (or conversely, learn a system of SDEs satisfying this), we may conclude that

μ is a stationary density of the solution $\{\mathbf{x}_t\}$ to the SDEs. This insight suggests that it is sufficient to find the function w achieving the *largest* deviation from $\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{A}h(\mathbf{x})] = 0$ among all test functions h. In the following, we derive a closed form for this maximum deviation over a sufficiently-rich, *infinite* set of functions as well as the witness function w achieving this maximum (or specifically, supremum). We sketch our proofs and defer their formal arguments to Appendix B.

3.1 Bounding the Deviation from Stationarity

Our key idea for bounding the functional in (5) is to consider functions h in an RKHS $\mathcal{H}$. We show that this allows us to derive a closed-form expression for the supremum of (5) over an expressive, infinite subset of functions in the RKHS. In the following, let $\mathcal{H}$ be the RKHS of a kernel k as introduced in Section 2, and let $\mathcal{F} := \{h \in \mathcal{H} : \|h\|_{\mathcal{H}} \leq 1\}$ be the unit ball of $\mathcal{H}$.

To begin, we first focus on the closely-related functional $\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}h(\mathbf{x})]$ involving the operator $\mathcal{L}$ instead of the generator $\mathcal{A}$. Recall that the operator $\mathcal{L}$ coincides with the generator $\mathcal{A}$ of the diffusion $\{\mathbf{x}_t\}$ solving the SDEs (2) when applied to the well-behaved functions C_c^2 (Section 2). For this functional, we can show that there exists a representer function $g_{\mu,\mathcal{L}}$ in the RKHS $\mathcal{H}$, whose inner product with any function $h \in \mathcal{H}$ allows evaluating the functional:

Lemma 1 *Let μ be a probability density over $\mathbb{R}^d$ and assume that the functions f, σ, and the partial[1] derivatives $\partial/\partial x_{i,i} k(\mathbf{x},\mathbf{x})$ and $\partial^2/\partial x_{i,i}\partial x_{j,j} k(\mathbf{x},\mathbf{x})$ are square-integrable under μ. Then, there exists a unique function $g_{\mu,\mathcal{L}} \in \mathcal{H}$ such that, for any $h \in \mathcal{H}$,*

$$\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}h(\mathbf{x})] = \langle h, g_{\mu,\mathcal{L}}\rangle_{\mathcal{H}}\,.$$

Moreover, $g_{\mu,\mathcal{L}}(\cdot) = \mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}_{\mathbf{x}} k(\mathbf{x},\cdot)]$. Here, $\mathcal{L}_{\mathbf{x}}$ indicates that $\mathcal{L}$ is applied to the argument $\mathbf{x}$.

The representation in Lemma 1 allows us to derive a closed form for the supremum of $\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}h(\mathbf{x})]$ over the unit ball $\mathcal{F}$, because the inner product with functions of $\mathcal{F}$ is maximized by the unit-norm function aligned with $g_{\mu,\mathcal{L}}$, that is, by $w_{\mu,\mathcal{L}} := g_{\mu,\mathcal{L}}/\|g_{\mu,\mathcal{L}}\|_{\mathcal{H}}$. Their inner product is then $\|g_{\mu,\mathcal{L}}\|_{\mathcal{H}}$. We will refer to the square of this RKHS norm as the *kernel deviation from stationarity* $\mathrm{KDS}(\mathcal{L},\mu;\mathcal{F})$:

Theorem 2 *Under the assumptions of Lemma 1, it holds that*

$$\sup_{h\in\mathcal{F}} \mathbb{E}_{\mathbf{x}\sim\mu}\big[\mathcal{L}h(\mathbf{x})\big] = \sqrt{\mathrm{KDS}(\mathcal{L},\mu;\mathcal{F})}\,,$$

where $\mathrm{KDS}(\mathcal{L},\mu;\mathcal{F}) := \mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}_{\mathbf{x}}\,\mathbb{E}_{\mathbf{x}'\sim\mu}[\mathcal{L}_{\mathbf{x}'} k(\mathbf{x},\mathbf{x}')]]$. Under additional regularity conditions on f,σ,k, and μ, we may interchange limits and write $\mathrm{KDS}(\mathcal{L},\mu;\mathcal{F}) = \mathbb{E}_{\mathbf{x}\sim\mu,\mathbf{x}'\sim\mu}\big[\mathcal{L}_{\mathbf{x}}\mathcal{L}_{\mathbf{x}'} k(\mathbf{x},\mathbf{x}')\big]$.

When thinking of $\mathcal{L}$ as the generator $\mathcal{A}$, the witness $w_{\mu,\mathcal{L}}$ is the smooth RKHS function that is subject to the largest infinitesimal change in the diffusion when evaluated in expectation over μ. Moreover, the KDS measures the maximal absolute deviation from (5) of any function in $\mathcal{F}$. More broadly, the KDS relates to (5) in the same way the maximum mean discrepancy (MMD, Gretton et al., 2012) relates to integral probability metrics (Müller, 1997), where the MMD is defined as $\mathrm{MMD}(\mu,\nu;\mathcal{F}) := \sup_{h\in\mathcal{F}} \mathbb{E}_{\mathbf{x}\sim\mu}[h(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\nu}[h(\mathbf{x})]$ for densities μ and ν. Both the KDS and the MMD express the maximum discrepancy between a target density μ and a model ($\mathcal{L}$ or ν, respectively) in a kernelized, closed form over $\mathcal{F}$. We leverage this learning perspective later for learning stationary SDEs from data, since the SDE functions f and σ enter the KDS via the operator $\mathcal{L}$.

3.2 Consistency

While the KDS measures a deviation from stationarity, it may not be consistent—$\mathrm{KDS}(\mathcal{L},\mu;\mathcal{F}) = 0$ may not guarantee that all SDEs entailing the operator $\mathcal{L}$ indeed induce the stationary density μ. Guaranteeing this requires that the equality of the functional of $\mathcal{A}$ in (5) holds for all functions in a core for $\mathcal{A}$. However, the SDE-parameterized operator $\mathcal{L}$ only coincides with the generator $\mathcal{A}$ of the diffusion for all $h \in C_c^2$ (Section 2). Moreover, $\mathcal{H}$ may not be dense in a core for $\mathcal{A}$ and thus fail to be sufficiently rich for testing the condition in (5).

To link the KDS to $\mathcal{A}$, we need to relate a core for $\mathcal{A}$ to the functions spanned by the RKHS $\mathcal{H}$. In general, the relationship between these two function spaces strongly depends on the generality of the

[1] Like Steinwart and Christmann (2008), we use $\partial/\partial x_{i,i}$ to denote the first-order partial derivative with respect to both function arguments, that is, $\partial/\partial x_{i,i} k(\mathbf{x},\mathbf{x}) := \partial/\partial u_i \partial/\partial v_i k(\mathbf{u},\mathbf{v})|_{\mathbf{u}:=\mathbf{x},\mathbf{v}:=\mathbf{x}}$.

functions f, σ defining the SDEs (Ethier and Kurtz, 1986, Chapter 8) and the kernel k (Christmann and Steinwart, 2010; Kanagawa et al., 2018). In the following, we show the consistency of the KDS for the Matérn kernel $k_{\nu,\gamma}$, which generalizes the Gaussian kernel $k_\gamma(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2/2\gamma^2)$ (Appendix A). We achieve this by showing that a core for $\mathcal{A}$ is dense in the Matérn RKHS with respect to a Sobolev norm. Building on this, we then prove that, for any h in the core, there always exists a nearby RKHS element ensuring that $\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{A}h(\mathbf{x})]$ is arbitrarily small:

Theorem 3 *Let $k_{\nu,\gamma}$ be a Matérn kernel with $\nu > 2$ defined over $\mathbb{R}^d$, and let $\mathcal{F}$ be the unit ball of its RKHS. Let μ be a probability density over $\mathbb{R}^d$ and f, σ be bounded functions with $\sigma\sigma^\top$ positive definite that define the SDEs in (2). Then, μ is a stationary density of the stochastic process $\{\mathbf{x}_t\}$ solving the SDEs if and only if*

$$\mathrm{KDS}(\mathcal{L}, \mu; \mathcal{F}) = 0\,.$$

3.3 The KDS as a Learning Objective

The KDS provides a closed-form expression for the maximum stationarity violation of any $h \in \mathcal{F}$. Since it quantifies this violation (as an RKHS norm), the KDS serves as an objective we can minimize to fit a system of SDEs to a target density μ. Specifically, given a dataset $D = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ of i.i.d. samples $\mathbf{x}^{(n)} \sim \mu$, we can compute the sample approximation of the $\mathrm{KDS}(\mathcal{L}, \mu; \mathcal{F})$ as

$$\hat{\mathrm{KDS}}(\mathcal{L}, D; k) := \frac{1}{N^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \mathcal{L}_{\mathbf{x}} \mathcal{L}_{\mathbf{x}'} k(\mathbf{x}^{(n)}, \mathbf{x}^{(m)})\,. \tag{6}$$

When the SDE model $f_{\boldsymbol{\theta}}, \sigma_{\boldsymbol{\theta}}$ is parameterized by some $\boldsymbol{\theta}$, we will indicate this in the operator $\mathcal{L}$ by a superscript (here as $\mathcal{L}^{\boldsymbol{\theta}}$). The KDS depends on the SDE parameters $\boldsymbol{\theta}$ through the operator $\mathcal{L}^{\boldsymbol{\theta}}$. Thus, minimizing the KDS enables us to estimate the parameters of a stochastic dynamical system without backpropagating gradients through time. The function $\mathcal{L}^{\boldsymbol{\theta}}_{\mathbf{x}} \mathcal{L}^{\boldsymbol{\theta}}_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}')$ inside the KDS is fully differentiable with respect to the model parameters $\boldsymbol{\theta}$. Notably, the KDS is exact up to the Monte Carlo approximation of the expectations over the target μ made in (6)—there are no SDE model components we need to sample from, roll out, reparameterize, or approximate. Appendix C provides an explicit form of (6) and an illustration of the empirical KDS and its gradients.

4 Stationary Diffusions as Causal Models

In this section, we describe how stationary diffusions can serve as causal models. To facilitate this exposition, we first leave the KDS aside and focus on discussing causality in SDEs, intervention models, and related properties. To conclude, we then leverage the KDS as an objective for learning stationary diffusions as causal models from a collection of interventional datasets.

4.1 Modeling Causal Dependencies with Stationary SDEs

Probabilistic causal models of a system $\mathbf{x} \in \mathbb{R}^d$ entail more than the *observational* density of the variables. A causal model contains additional information that characterizes the *interventional* densities of the system under interventions on its data-generating process (Peters et al., 2017). This information may be in the form of, say, functions f_j that explicitly relate the densities of x_j and remain invariant under interventions elsewhere, as in SCMs. Which causal model of a system is adequate depends on the application and the level of modeling granularity (Rubenstein et al., 2017; Schölkopf, 2022).

In this work, we propose to model the causal effects of the variables $\mathbf{x}$ via a stationary dynamical system of $\mathbf{x}$ over time t. Specifically, we model the time evolution of $\mathbf{x}_t \in \mathbb{R}^d$ with the stationary SDEs

$$\mathrm{d}\mathbf{x}_t = f_{\boldsymbol{\theta}}(\mathbf{x}_t)\mathrm{d}t + \sigma_{\boldsymbol{\theta}}(\mathbf{x}_t)\mathrm{d}\mathbb{W}_t\,, \tag{7}$$

with parameters $\boldsymbol{\theta} \in \mathbb{R}^k$ and observational stationary density $\mu(\mathbf{x})$. Time remains internal to the model—only the stationary densities of the system, which are time-invariant, form the probabilistic causal model of $\mathbf{x}$ and characterize its behavior under interventions. Our core idea is that the explicit time dimension enables propagating feedback cycles in the causal dependencies of the variables. By contrast, SCMs do not allow for cycles in the causal structure except under restrictive model and invertibility assumptions (see *Related Work* in Section 5).

Similar to the structural equations in SCMs, the differential equations in SDEs provide a *mechanistic* (or functional) model of the causal dependencies among the variables $\mathbf{x}$ (Peters et al., 2017; Schölkopf,

2022). The causal mechanisms f_j and σ_j model which variables in $\mathbf{x}$ affect the variable x_j via an explicit functional dependency that holds independent of perturbations of the variables or the functions governing the other variables. When the mechanisms are independent, the SDEs factorize as

$$\mathrm{d}x_{jt} = f_{\boldsymbol{\theta}_j}(\mathbf{x}_t)_j \,\mathrm{d}t + \sigma_{\boldsymbol{\theta}_j}(\mathbf{x}_t)_j \,\mathrm{d}\mathbb{W}_t \,, \tag{8}$$

with $f_j : \mathbb{R}^d \to \mathbb{R}$ and $\sigma_j : \mathbb{R}^d \to \mathbb{R}^b$ not sharing any parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d\}$. Ultimately, both SCMs and stationary SDEs should be thought of as different abstractions of the physical processes underlying our measurements $\mathbf{x} \in \mathbb{R}^d$ (e.g., Peters et al., 2017, Section 2.3.3), with SDEs characterizing the processes explicitly over time.

4.2 Intervention Models

Interventions modifying the data-generating process of SDEs can be modeled in various ways and often in analogy to SCMs (Eberhardt and Scheines, 2007). We formalize an intervention by a model with parameters $\boldsymbol{\phi} \in \mathbb{R}^k$ that characterizes its effect on the SDEs. The intervention $\boldsymbol{\phi}$ transforms $f_{\boldsymbol{\theta}}$ and $\sigma_{\boldsymbol{\theta}}$ into the modified mechanisms $f_{\boldsymbol{\theta},\boldsymbol{\phi}}$ and $\sigma_{\boldsymbol{\theta},\boldsymbol{\phi}}$ such that the system in (7) now evolves as

$$\mathrm{d}\mathbf{x}_t = f_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}_t)\mathrm{d}t + \sigma_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}_t)\mathrm{d}\mathbb{W}_t \,. \tag{9}$$

The interventional density $\mu_{\boldsymbol{\phi}}(\mathbf{x})$ denotes the stationary density of the modified SDEs in (9). For example, some real-world perturbations may be modeled as shift-scale interventions, in which the mechanisms $f_{\boldsymbol{\theta}}$ and $\sigma_{\boldsymbol{\theta}}$ of a variable x_j are shifted and scaled by some scalars δ, γ, respectively, as

$$f_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x})_j = f_{\boldsymbol{\theta}}(\mathbf{x})_j + \delta \qquad \text{and} \qquad \sigma_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x})_j = \gamma\, \sigma_{\boldsymbol{\theta}}(\mathbf{x})_j \tag{10}$$

where $\boldsymbol{\phi} = \{\delta, \gamma\}$. Analogous shift interventions have been studied in acyclic and cyclic SCMs (Zhang et al., 2021; Rothenhäusler et al., 2015).

4.3 Properties

Complexity Stationary diffusions can be modeled by arbitrary functions f, σ. Even with $\sigma = \sqrt{2}\mathbf{I}$, they can characterize any observational density μ via its score function $f = -\nabla_{\mathbf{x}} \log \mu$ (as a Langevin diffusion). When σ is non-diagonal, the driving noise of the equations $\mathrm{d}\mathbf{x}_t$ becomes correlated, which can model confounding. Thus, the function classes of f and σ determine the complexity of the densities modeled by the diffusion, not $\{\mathbb{W}_t\}$ alone. Besides some notable exceptions (Immer et al., 2023), the assumptions of stationary diffusions are less restrictive than those of SCMs, where the noise defines the distributional family *a priori*.

Stability Using diffusions for causal modeling relies on the stationarity, i.e., stability, of the SDEs. For general $f_{\boldsymbol{\theta}}$ and $\sigma_{\boldsymbol{\theta}}$, stability is not guaranteed, particularly when randomly initializing the model parameters $\boldsymbol{\theta}$. For example, in linear systems $\mathrm{d}\mathbf{x}_t = (\mathbf{a} + \mathbf{B}\mathbf{x}_t)\mathrm{d}t + \mathbf{C}\mathrm{d}\mathbb{W}_t$, stability requires that the eigenvalues of $\mathbf{B}$ have negative real parts (Särkkä and Solin, 2019). Guaranteeing stability under interventions, however, is possible in certain cases: in linear systems, the shift-scale interventions in (10) do not affect stability. More generally, Theorem 3 shows that $\mathrm{KDS} = 0$ can guarantee stability and act as a certificate, even for complex model classes.

Identifiability Causal modeling aims at generalizing to (combinations of) intervention classes when learning a model from a set of observed interventions. Generalizing to unseen perturbations may not require fully identifying $\boldsymbol{\theta}$. For SDEs in particular, a density μ does not uniquely identify the true parameters $\boldsymbol{\theta}$ in a model class without unverifiable assumptions: changing the *speed* of a diffusion via $\mathrm{d}\mathbf{x}_t = s f(\mathbf{x}_t)\mathrm{d}t + \sqrt{s}\,\sigma(\mathbf{x}_t)\mathrm{d}\mathbb{W}_t$ for $s > 0$ leaves the stationary density unchanged. The operator $s\mathcal{L}$ satisfies the same stationarity conditions as $\mathcal{L}$ (Hansen and Scheinkman, 1995). While linear systems are identifiable up to speed scaling under specific sparsity conditions (Dettling et al., 2022), it is, to our knowledge, not yet known to what degree multiple interventional densities $\mu_{\boldsymbol{\phi}}$ identify stationary SDEs. As we investigate in Section 6, stationary diffusions empirically allow generalizing to unseen interventions, hence weaker notions of identifiability may be appropriate.

4.4 Learning Stationary Diffusions from Interventional Data

We can use the KDS derived in Section 3 as an objective for learning a causal stationary diffusion model from a collection of interventional datasets. We consider the setting in which a system of stationary SDEs $f_{\boldsymbol{\theta}^*}, \sigma_{\boldsymbol{\theta}^*}$ is perturbed by some interventions $\boldsymbol{\phi}^*_{1:m} = \{\boldsymbol{\phi}^*_1, \ldots, \boldsymbol{\phi}^*_m\}$, whose parameters may be unknown. The observations consist of m corresponding datasets D_i with $\mathbf{x} \sim \mu_{\boldsymbol{\phi}^*_i}$ for

each $\mathbf{x} \in D_i$. Our goal is to learn a stationary SDE model that is jointly consistent with the observed interventions, i.e., induces the observed stationary densities under the considered intervention class.

To infer the parameters $\boldsymbol{\theta}$, we optimize $\boldsymbol{\theta}$ such that the interventional densities induced under $\boldsymbol{\theta}$ and $\boldsymbol{\phi}^*_{1:m}$ fit the observed distributions $\mu_{\boldsymbol{\phi}^*_i}$. When the model modifications $\boldsymbol{\phi}^*_{1:m}$ are unknown, which is often the case in practice, we learn $\boldsymbol{\phi}_{1:m}$ alongside $\boldsymbol{\theta}$. Observing *multiple* interventions makes this joint inference problem well-posed, in particular when the interventions are sparse, since $\boldsymbol{\theta}$ is shared for all interventions (Schölkopf, 2022). Using the KDS, the model $\boldsymbol{\theta}$ and the interventions $\boldsymbol{\phi}_{1:m}$ can be learned with gradient descent. At each iteration, we draw a batch from a dataset D_i and update $\boldsymbol{\theta}$ (and $\boldsymbol{\phi}_i$) using the KDS gradients of the intervened SDEs $f_{\boldsymbol{\theta},\boldsymbol{\phi}_i}$ and $\sigma_{\boldsymbol{\theta},\boldsymbol{\phi}_i}$. To mitigate overfitting, we apply a *group lasso* penalty $R(\boldsymbol{\theta}_j)$ separately to each $\boldsymbol{\theta}_j$ to encourage sparse dependencies on the other variables (Yuan and Lin, 2006). Overall, the optimizer steps for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}_i$ are proportional to

$$\propto -\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}_i} \left(\hat{\mathrm{KDS}}(\mathcal{L}^{\boldsymbol{\theta},\boldsymbol{\phi}_i}, D; k) + \lambda \sum_{j=1}^{d} R(\boldsymbol{\theta}_j) \right), \tag{11}$$

with $\lambda > 0$. When learning both $f_{\boldsymbol{\theta}}$ and $\sigma_{\boldsymbol{\theta}}$, the invariance to speed scaling described in Section 4.3 can cause an instability close to convergence, as decreasing the speed s via $sf(\mathbf{x})$ and $\sqrt{s}\sigma(\mathbf{x})$ shrinks the KDS. This can be prevented by fixing (the scale of) subsets of the parameters of f or σ, for example, the self-regulating dependence of $f_j(\mathbf{x})$ on x_j. Empirically, minimizing the KDS was sufficient in combination with sparsity regularization to ensuring stability upon convergence.

5 Related Work

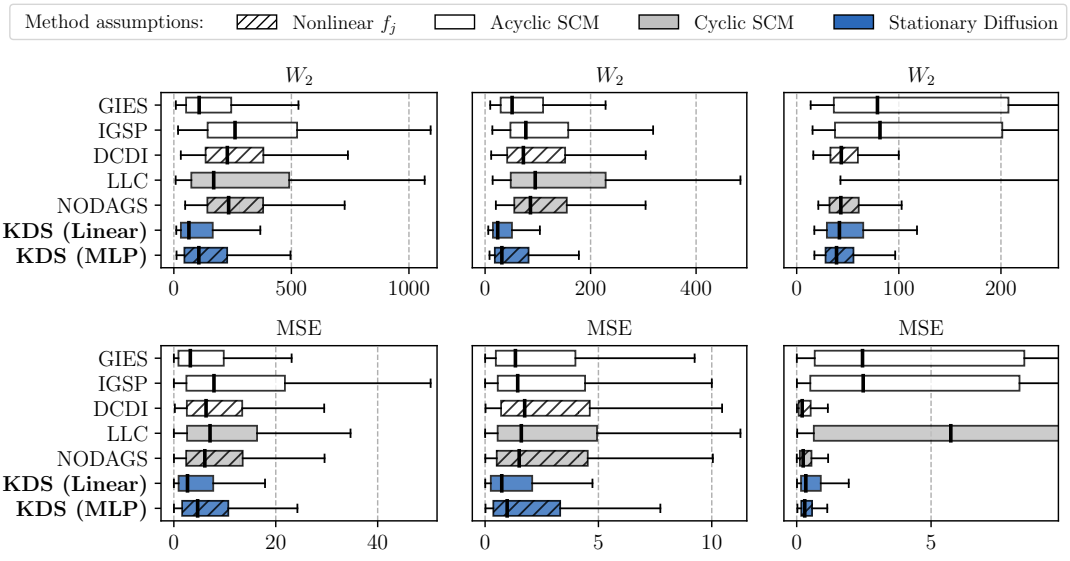
Causality in dynamical systems When observing dynamical systems over time, fields like Granger causality (Granger, 1969), autoregressive modeling (Hyvärinen et al., 2010), and system identification (Ljung, 1998) allow inferring notions of causation. Hansen and Sokol (2014) and Peters et al. (2022) formally study interventions in SDE systems observed over time. Contrary to these time series settings, we adopt the novel perspective of using the stationary distributions of SDEs to learn a time-independent causal model. Our approach makes explicit that causal models, including SCMs, are abstractions of processes taking place in time (e.g., Peters et al., 2017, Section 2.3.3)—even when causation occurs on scales that either are not or cannot be measured as time series. Varando and Hansen (2020) also study stationary SDEs in the linear case, but they interpret them as probabilistic graphical models via the Lyapunov equation, not considering causality or interventions. Mooij et al. (2013) and Bongers et al. (2022) investigate how equilibria of differential equations relate to SCMs.

Cyclic graphical modeling Several works propose to interpret SCMs in ways that enable learning cycles (Richardson, 1996; Lacerda et al., 2008; Mooij et al., 2011; Hyttinen et al., 2012; Mooij and Heskes, 2013; Rothenhäusler et al., 2015; Sethuraman et al., 2023). These approaches usually assume additive noise and linearity, sometimes with restrictions on the feedback, and require a unique solution $\mathbf{x}$ to the system of equations $\mathbf{x} = f(\mathbf{x}) + \boldsymbol{\epsilon}$ given any possible $\boldsymbol{\epsilon}$ (Bongers et al., 2021). Our proposal of modeling causality with stationary SDEs shares the intuition of an equilibrium, but it expands on the insight that cyclicity necessarily introduces a notion of time, ultimately enabling us to drop prior model restrictions. As real-world processes evolve in time, some challenge the notion of aggregating causality in graphical models altogether (Dawid, 2010; Aalen et al., 2016).

Statistical inference and kernels The idea of producing diffusions that imply certain densities goes back to Wong (1964), who linked SDEs with polynomial functions f and σ to the Pearson distributions. In econometrics, the infinitesimal generator and Equation (5) are known tools for fitting diffusion models, but usually with specific parameterizations and test functions (see Aït-Sahalia et al., 2010, Section 3, for an overview). The KDS extends these works by introducing a general-purpose characterization of stationarity that covers an infinite class of test functions in closed form. Our techniques establish novel connections between SDEs and RKHSs and build on kernel properties previously used by, for example, kernel mean embeddings (Smola et al., 2007), the MMD (Gretton et al., 2012), and the kernelized Stein discrepancy (Liu et al., 2016).

6 Experiments

The downstream purpose of causal modeling is to predict the effects of interventions in a system. To evaluate this, we compare the interventional densities predicted by stationary diffusions to those by



(a) **Cyclic Linear SCMs** (b) **Cyclic Linear SDEs** (c) **SERGIO**

Figure 2: Benchmarking results ($d = 20$ variables, Erdős-Rényi sparsity structure). Metrics are computed from 10 test interventions on unseen target variables in 50 randomly-generated systems. Box plots show medians and interquartile ranges (IQR). Whiskers extend to the largest value inside 1.5 times the IQR length from the boxes. Overall, stationary diffusions are the most accurate at predicting the effects of interventions on unseen targets, measured in terms of both W_2 (↓) and MSE (↓).

existing approaches. All methods first learn a causal model from interventional data with known target variables and then predict the distributions resulting from unseen interventions by sampling from the learned models. The test interventions are out-of-distribution, that is, on unseen targets.

Data We evaluate the methods on sparse cyclic linear systems (SCMs and stationary SDEs) and expression data of sparse gene regulatory networks. For the latter, we simulate the SERGIO model by Dibaeinia and Sinha (2020), which requires acyclic dependencies, without technical noise. For each system, we sample observational data and interventional data for 10 train and 10 test interventions on disjoint variables, each dataset containing 1000 observations. In the linear systems, we perform shift interventions; in SERGIO, we implement overexpression gene perturbations (e.g., Norman et al., 2019). All datasets are standardized by the mean and variance of the observational data.

Models We learn stationary diffusions with linear and MLP mechanisms $f(\mathbf{x})$ and a constant matrix $\sigma(\mathbf{x}) = \mathrm{diag}(\boldsymbol{\sigma})$, $\boldsymbol{\sigma} \in \mathbb{R}^d$. Their model definition and group lasso regularizers are given in Appendix D. To estimate the KDS, we use the Gaussian kernel $k_\gamma(\mathbf{x}, \mathbf{x}')$. We compare with five SCM approaches that learn from interventional data: GIES (Hauser and Bühlmann, 2012), IGSP (Wang et al., 2017), and DCDI (Brouillard et al., 2020). We also benchmark LLC (Hyttinen et al., 2012) and NODAGS (Sethuraman et al., 2023), both of which allow modeling cycles.

Metrics The test interventions performed to query the learned causal models are shift interventions that match the interventional mean of the target variable in the held-out data. To allow comparing methods with explicit and implicit densities, we report the Wasserstein distance W_2 between the true and predicted interventional data. We also report the mean squared error (MSE) of the true and predicted empirical means (Zhang et al., 2022).

Results Figures 2a and 2b present the results for the cyclic linear SCM and stationary SDE systems, respectively. Both the linear and MLP diffusions learned via the KDS achieve the most accurate interventional density predictions in both the W_2 and MSE metrics. The acyclic approaches, in particular GIES, show competitive performance, highlighting a trade-off between model complexity and the entailed inference challenge, even when the data qualitatively violates acyclicity. In contrast, the cyclic SCM approaches underperform, particularly LLC, whose model assumptions—apart from data standardization—perfectly align with this setting. The synthetic gene expression data assesses all methods under model mismatch. Figure 2c shows that stationary diffusions, especially the nonlinear MLP diffusion, match the best baselines DCDI and NODAGS, which also model nonlinearity.

Acknowledgments

Many thanks to Ya-Ping Hsieh and Mohammad Reza Karimi for the engaging discussions on SDEs in the early stages of this work. We additionally thank Charlotte Bunne, Paweł Czyż, Jonas Rothfuss, and Scott Sussex for their helpful comments on different versions of the manuscript. This work has also greatly benefited from exchanges on kernels with Parnian Kassraie, Mojmír Mutný, Jonas Rothfuss, and Ingo Steinwart; technical feedback by Zebang Shen; and discussions on connections to Stein's method with Jonas Hübotter, for which we are very thankful.

This research was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program grant agreement no. 815943 and the Swiss National Science Foundation under NCCR Automation, grant agreement 51NF40 180545.

References

Aalen, O. O., Røysland, K., Gran, J. M., Kouyos, R., and Lange, T. (2016). Can we believe the DAGs? A comment on the relationship between causal DAGs and mechanisms. *Statistical methods in medical research*, 25(5):2294–2314.

Adams, R. A. and Fournier, J. J. (2003). *Sobolev spaces*. Elsevier, 2nd edition.

Aït-Sahalia, Y., Hansen, L. P., and Scheinkman, J. A. (2010). Operator methods for continuous-time Markov processes. *Handbook of financial econometrics: tools and techniques*, pages 1–66.

Bongers, S., Blom, T., and Mooij, J. M. (2022). Causal modeling of dynamical systems. *arXiv preprint arXiv:1803.08784*.

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. (2021). Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. (2018). JAX: composable transformations of Python+NumPy programs.

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. (2020). Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877.

Bunne, C., Papaxanthos, L., Krause, A., and Cuturi, M. (2022). Proximal optimal transport modeling of population dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 6511–6528. PMLR.

Christmann, A. and Steinwart, I. (2010). Universal kernels on non-standard input spaces. *Advances in neural information processing systems*, 23.

Cuturi, M., Meng-Papaxanthos, L., Tian, Y., Bunne, C., Davis, G., and Teboul, O. (2022). Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*.

Dawid, A. P. (2010). Beware of the DAG! In *Causality: objectives and assessment*, pages 59–86. PMLR.

Dettling, P., Homs, R., Améndola, C., Drton, M., and Hansen, N. R. (2022). Identifiability in continuous Lyapunov models. *arXiv preprint arXiv:2209.03835*.

Dibaeinia, P. and Sinha, S. (2020). SERGIO: a single-cell expression simulator guided by gene regulatory networks. *Cell systems*, 11(3):252–271.

Eberhardt, F. and Scheines, R. (2007). Interventions and causal inference. *Philosophy of science*, 74(5):981–995.

Ethier, S. N. and Kurtz, T. G. (1986). *Markov processes: characterization and convergence*. John Wiley & Sons.

Genevay, A., Peyré, G., and Cuturi, M. (2018). Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Hansen, L. P. and Scheinkman, J. A. (1995). Back to the future: Generating moment implications for continuous-time Markov processes. *Econometrica*, 63(4):767–804.

Hansen, N. and Sokol, A. (2014). Causal interpretation of stochastic differential equations. *Electronic Journal of Probability*, 19:1–24.

Hauser, A. and Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1):2409–2464.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. (2012). Learning linear cyclic causal models with latent variables. *The Journal of Machine Learning Research*, 13(1):3387–3439.

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. (2010). Estimation of a structural vector autoregression model using non-Gaussianity. *Journal of Machine Learning Research*, 11(5).

Immer, A., Schultheiss, C., Vogt, J. E., Schölkopf, B., Bühlmann, P., and Marx, A. (2023). On the identifiability and estimation of causal location-scale noise models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 14316–14332. PMLR.

Jacobsen, M. (1993). A brief account of the theory of homogeneous Gaussian diffusions in finite dimensions. *Frontiers in Pure and Applied Probability 1*, pages 86–94.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018). Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*.

Lacerda, G., Spirtes, P., Ramsey, J., and Hoyer, P. O. (2008). Discovering cyclic causal models by independent components analysis. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, page 366–374, Arlington, Virginia, USA. AUAI Press.

Liu, Q., Lee, J., and Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR.

Ljung, L. (1998). *System identification*. Springer.

Mooij, J. M. and Heskes, T. (2013). Cyclic causal discovery from continuous equilibrium data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 431–439, Arlington, Virginia, USA. AUAI Press.

Mooij, J. M., Janzing, D., Heskes, T., and Schölkopf, B. (2011). On causal discovery with cyclic additive noise models. *Advances in neural information processing systems*, 24.

Mooij, J. M., Janzing, D., and Schölkopf, B. (2013). From ordinary differential equations to structural causal models: The deterministic case. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 440–448, Arlington, Virginia, USA. AUAI Press.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443.

Norman, T. M., Horlbeck, M. A., Replogle, J. M., Ge, A. Y., Xu, A., Jost, M., Gilbert, L. A., and Weissman, J. S. (2019). Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793.

Øksendal, B. (2003). *Stochastic differential equations*. Springer.

Pearl, J. (2009). *Causality*. Cambridge university press.

Peters, J., Bauer, S., and Pfister, N. (2022). Causal models for dynamical systems. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 671–690. Association for Computing Machinery.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*, volume 1. Springer.

Richardson, T. (1996). A polynomial-time algorithm for deciding Markov equivalence of directed cyclic graphical models. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, page 462–469, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Rothenhäusler, D., Heinze, C., Peters, J., and Meinshausen, N. (2015). backShift: Learning causal cyclic graphs from unknown shift interventions. *Advances in Neural Information Processing Systems*, 28.

Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J. M., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. (2017). Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, page ID 11.

Särkkä, S. and Solin, A. (2019). *Applied stochastic differential equations*, volume 10. Cambridge University Press.

Schölkopf, B. (2022). Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 765–804. Association for Computing Machinery.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels*. MIT press.

Sethuraman, M. G., Lopez, R., Mohan, R., Fekri, F., Biancalani, T., and Hütter, J.-C. (2023). NODAGS-Flow: Nonlinear cyclic causal structure learning. In *International Conference on Artificial Intelligence and Statistics*, pages 6371–6387. PMLR.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10).

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A Hilbert space embedding for distributions. In *Algorithmic Learning Theory: 18th International Conference, ALT 2007, Sendai, Japan, October 1-4, 2007. Proceedings 18*, pages 13–31. Springer.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Varando, G. and Hansen, N. R. (2020). Graphical continuous Lyapunov models. In *Conference on Uncertainty in Artificial Intelligence*, pages 989–998. PMLR.

Wang, Y., Solus, L., Yang, K., and Uhler, C. (2017). Permutation-based causal inference algorithms with interventions. *Advances in Neural Information Processing Systems*, 30.

Wendland, H. (2004). *Scattered data approximation*, volume 17. Cambridge university press.

Wong, E. (1964). The construction of a class of stationary Markoff processes. In *Proc. Sympos. Appl. Math., Vol. XVI*, pages 264–276. Amer. Math. Soc., Providence, RI.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67.

Zhang, J., Cammarata, L., Squires, C., Sapsis, T. P., and Uhler, C. (2022). Active learning for optimal intervention design in causal models. *arXiv preprint arXiv:2209.04744*.

Zhang, J., Squires, C., and Uhler, C. (2021). Matching a desired causal state via shift interventions. *Advances in Neural Information Processing Systems*, 34:19923–19934.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). DAGs with NO TEARS: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31.

A Additional Background

A.1 Euler-Maruyama Method

To approximate the solutions to SDEs, we use the Euler-Maruyama method (Särkkä and Solin, 2019, Section 8.2). The Euler-Maruyama approximation of sample paths of the diffusion solving (2) is given by

$$\mathbf{x}_{l+1} := \mathbf{x}_l + f(\mathbf{x}_l)\Delta t + \sigma(\mathbf{x}_l)\boldsymbol{\xi}_l\sqrt{\Delta t} \tag{12}$$

for some step size Δt and independent vectors $\boldsymbol{\xi}_l \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. To generate L samples from the stationary density $\mu(\mathbf{x})$ of (2), we simulate a single sample path and then select every k-th state $\mathbf{x}_{l \cdot k}$ for $l \in \{1, \ldots, L\}$ as a sample, where k is a thinning factor as in Markov chain Monte Carlo. In our experiments, we sample $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and use a step size of $\Delta t = 0.01$, a thinning factor of 500, and 100 samples of burn-in, which we chose based on autocorrelation diagnostics of the thinned Markov chains.

A.2 Sobolev Spaces

Some of our theoretical results build on the notion of Sobolev spaces. While not required here, we recommend Adams and Fournier (2003) for a detailed introduction. The Sobolev norm $\|\cdot\|_{m,p}$ of a function f sums the L_p norms of all its partial derivatives up to order m and is defined as

$$\|f\|_{m,p} := \left(\sum_{\mathbf{n} \in \mathbb{N}_0^d : |\mathbf{n}| \leq m} \left\| \frac{\partial^{n_1}}{\partial x_1^{n_1}} \cdots \frac{\partial^{n_d}}{\partial x_d^{n_d}} f \right\|_p^p \right)^{1/p}$$

for $1 \leq p < \infty$. Here, $\|\cdot\|_p$ is the L_p norm defined as $\|f\|_p = \left(\int_{\mathbb{R}^d} |f(\mathbf{x})|^p \mathrm{d}\mathbf{x}\right)^{1/p}$. The Sobolev space $W^{m,p}$ contains all functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\|f\|_{m,p} < \infty$. Moreover, the space $W_c^{m,p}$ is defined as the closure of C_c^∞ in $W^{m,p}$ (Adams and Fournier, 2003, Section 3.2).

A.3 Matérn Kernel

The Matérn kernel $k_{\nu,\gamma}$ with smoothness and scale parameters $\nu, \sigma > 0$ can be seen as a generalization of the Gaussian kernel that allows controlling the smoothness of the RKHS functions. We write the Matérn kernel $k_{\nu,\gamma}(\mathbf{x}, \mathbf{x}')$ in terms of the distance $r = \|\mathbf{x} - \mathbf{x}'\|_2$ as

$$k_{\nu,\gamma}(r) := \tfrac{2^{1-\nu}}{\Gamma(\nu)} \left(\tfrac{\sqrt{2\nu}\, r}{\gamma} \right)^\nu K_\nu \left(\tfrac{\sqrt{2\nu}\, r}{\gamma} \right), \tag{13}$$

where Γ is the gamma function and K_ν is a modified Bessel function of the second kind and order ν (Rasmussen and Williams, 2006, Equation 4.14). Common special cases of $k_{\nu,\gamma}$ have the following explicit forms:

$$\begin{aligned}
k_{\nu=1/2,\gamma}(r) &= \exp\left(-\tfrac{r}{\gamma}\right) \\
k_{\nu=3/2,\gamma}(r) &= \left(1 + \tfrac{\sqrt{3}r}{\gamma}\right) \exp\left(-\tfrac{\sqrt{3}r}{\gamma}\right) \\
k_{\nu=5/2,\gamma}(r) &= \left(1 + \tfrac{\sqrt{5}r}{\gamma} + \tfrac{5r^2}{3\gamma^2}\right) \exp\left(-\tfrac{\sqrt{5}r}{\gamma}\right)
\end{aligned}$$

The Gaussian kernel $k_\gamma(r) = \exp(-r^2/2\gamma^2)$ is obtained from $k_{\nu,\gamma}$ as $\nu \to \infty$.

The following two results will be useful for proving Theorem 3. The first statement was originally shown by Wendland (2004, Corollary 10.48) and follows from Rasmussen and Williams (2006, Equation 4.15), linking the Matérn RKHS to the Sobolev spaces. The second result concerns the differentiability of the Matérn kernel function:

Lemma 4 (Kanagawa et al., 2018, Example 2.8) *The RKHS $\mathcal{H}$ of a Matérn kernel $k_{\nu,\gamma}$ is norm-equivalent to the Sobolev space $W^{\nu+d/2,2}$. Specifically, we have $h \in \mathcal{H}$ if and only if $h \in W^{\nu+d/2,2}$. Moreover, there exist constants c_1, c_2 such that $c_1\|h\|_{\nu+d/2,2} \leq \|h\|_{\mathcal{H}} \leq c_2\|h\|_{\nu+d/2,2}$ for all $h \in \mathcal{H}$.*

Lemma 5 (Stein, 1999, Section 2.7, p. 32) *The Matérn covariance function $k_{\nu,\gamma}(r)$ is $2k$-times differentiable if and only if $\nu > k$.*

B Proofs

B.1 Proof of Lemma 1

Let $\mathcal{B}_{\mu,\mathcal{L}} : \mathcal{H} \to \mathbb{R}$ be the composition of the two operators $\mathbb{E}_{\mathbf{x}\sim\mu}$ and $\mathcal{L}$ defined as $\mathcal{B}_{\mu,\mathcal{L}}h := \mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}h(\mathbf{x})]$ for all $h \in \mathcal{H}$. The functional $\mathcal{B}_{\mu,\mathcal{L}}$ is linear since both the expectation and $\mathcal{L}$ are linear operators. Moreover, $\mathcal{B}_{\mu,\mathcal{L}}$ is continuous because the functional is bounded on the unit ball of functions in $\mathcal{H}$:

$$
\begin{aligned}
&|\mathcal{B}_{\mu,\mathcal{L}}h| \\
&= |\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}h(\mathbf{x})]| \\
&\leq \mathbb{E}_{\mathbf{x}\sim\mu}[|\mathcal{L}h(\mathbf{x})|] && \text{Jensen's inequality} \\
&= \mathbb{E}_{\mathbf{x}\sim\mu}\left[\left|f(\mathbf{x})\cdot\nabla_{\mathbf{x}}h(\mathbf{x}) + \tfrac{1}{2}\operatorname{tr}(\sigma(\mathbf{x})\sigma(\mathbf{x})^\top\nabla_{\mathbf{x}}\nabla_{\mathbf{x}}h(\mathbf{x}))\right|\right] && \text{by definition} \\
&\leq \mathbb{E}_{\mathbf{x}\sim\mu}\left[\left|f(\mathbf{x})\cdot\nabla_{\mathbf{x}}h(\mathbf{x})\right| + \tfrac{1}{2}\left|\operatorname{tr}(\sigma(\mathbf{x})\sigma(\mathbf{x})^\top\nabla_{\mathbf{x}}\nabla_{\mathbf{x}}h(\mathbf{x}))\right|\right] && \text{triangle inequality} \\
&\leq \mathbb{E}_{\mathbf{x}\sim\mu}\left[\|f(\mathbf{x})\|_2\|\nabla_{\mathbf{x}}h(\mathbf{x})\|_2 + \tfrac{1}{2}\|\sigma(\mathbf{x})\sigma(\mathbf{x})^\top\|_{\mathrm{F}}\|\nabla_{\mathbf{x}}\nabla_{\mathbf{x}}h(\mathbf{x})\|_{\mathrm{F}}\right] && \text{Cauchy-Schwarz} \\
&= \mathbb{E}_{\mathbf{x}\sim\mu}\left[\|f(\mathbf{x})\|_2\Big(\sum_{i=1}^{d}\left|\tfrac{\partial}{\partial x_i}h(\mathbf{x})\right|^2\Big)^{1/2} + \tfrac{1}{2}\|\sigma(\mathbf{x})\sigma(\mathbf{x})^\top\|_{\mathrm{F}}\Big(\sum_{i=1}^{d}\sum_{j=1}^{d}\left|\tfrac{\partial^2}{\partial x_i\partial x_j}h(\mathbf{x})\right|^2\Big)^{1/2}\right] \\
&\leq \|h\|_{\mathcal{H}}\cdot\mathbb{E}_{\mathbf{x}\sim\mu}\left[\|f(\mathbf{x})\|_2\Big(\sum_{i=1}^{d}\tfrac{\partial}{\partial x_{i,i}}k(\mathbf{x},\mathbf{x})\Big)^{1/2} + \tfrac{1}{2}\|\sigma(\mathbf{x})\sigma(\mathbf{x})^\top\|_{\mathrm{F}}\Big(\sum_{i=1}^{d}\sum_{j=1}^{d}\tfrac{\partial^2}{\partial x_{i,i}\partial x_{j,j}}k(\mathbf{x},\mathbf{x})\Big)^{1/2}\right]. \quad (*)
\end{aligned}
$$

The last inequality follows from the fact that the partial derivatives of RKHS functions $h \in \mathcal{H}$ are bounded as

$$
\left|\tfrac{\partial}{\partial x_i}h(\mathbf{x})\right| \leq \|h\|_{\mathcal{H}}\cdot\left(\tfrac{\partial}{\partial x_{i,i}}k(\mathbf{x},\mathbf{x})\right)^{1/2} \qquad \text{and} \qquad \left|\tfrac{\partial^2}{\partial x_i\partial x_j}h(\mathbf{x})\right| \leq \|h\|_{\mathcal{H}}\cdot\left(\tfrac{\partial^2}{\partial x_{i,i}\partial x_{j,j}}k(\mathbf{x},\mathbf{x})\right)^{1/2}
$$

(Steinwart and Christmann, 2008, Corollary 4.36). The notation $\partial/\partial x_{i,i}$ was defined in Footnote 1. By assumption, the squares of f, σ, and the partial derivatives $\partial/\partial x_{i,i}k(\mathbf{x},\mathbf{x})$ and $\partial^2/\partial x_{i,i}\partial x_{j,j}k(\mathbf{x},\mathbf{x})$ are square-integrable with respect to μ, hence the expectation in $(*)$ is as well. Thus, when applied to the unit ball of $\mathcal{H}$, for which $\|h\|_{\mathcal{H}} \leq 1$, the norm of $\mathcal{B}_{\mu,\mathcal{L}}$ is bounded. Hence, $\mathcal{B}_{\mu,\mathcal{L}}$ is a continuous linear functional, and by the Riesz representation theorem, there exists a unique $g_{\mu,\mathcal{L}} \in \mathcal{H}$ such that

$$
\mathcal{B}_{\mu,\mathcal{L}}h = \mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}_{\mathbf{x}}h(\mathbf{x})] = \langle h, g_{\mu,\mathcal{L}}\rangle_{\mathcal{H}}
$$

for all $h \in \mathcal{H}$. We obtain the explicit form for $g_{\mu,\mathcal{L}}$ by substituting $k(\,\cdot\,,\mathbf{x}') \in \mathcal{H}$ for h, which yields

$$
\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}_{\mathbf{x}}k(\mathbf{x},\mathbf{x}')] = \langle k(\,\cdot\,,\mathbf{x}'), g_{\mu,\mathcal{L}}\rangle_{\mathcal{H}}\,,
$$

where the notation $\mathcal{L}_{\mathbf{x}}$ makes explicit that the operator $\mathcal{L}$ is applied to the argument $\mathbf{x}$. By the reproducing property, we have $\langle k(\,\cdot\,,\mathbf{x}'), g_{\mu,\mathcal{L}}\rangle_{\mathcal{H}} = g_{\mu,\mathcal{L}}(\mathbf{x}')$ (Schölkopf and Smola, 2002). Hence, $g_{\mu,\mathcal{L}}(\mathbf{x}') = \mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}_{\mathbf{x}}k(\mathbf{x},\mathbf{x}')]$, and we can read off that $g_{\mu,\mathcal{L}}(\cdot) = \mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}_{\mathbf{x}}k(\mathbf{x},\,\cdot\,)]$.

■

B.2 Proof of Theorem 2

Using the representer function $g_{\mu,\mathcal{L}}$ in Lemma 1, we can express the supremum over $\mathcal{F}$ as

$$
\sup_{h\in\mathcal{F}}\mathbb{E}_{\mathbf{x}\sim\mu}\big[\mathcal{L}h(\mathbf{x})\big] = \sup_{h\in\mathcal{F}}\langle h, g_{\mu,\mathcal{L}}\rangle_{\mathcal{H}} = \Big\langle \tfrac{g_{\mu,\mathcal{L}}}{\|g_{\mu,\mathcal{L}}\|_{\mathcal{H}}}, g_{\mu,\mathcal{L}}\Big\rangle_{\mathcal{H}} = \|g_{\mu,\mathcal{L}}\|_{\mathcal{H}}\,.
$$

In the above, we used the fact that the norm $\langle h, g_{\mu,\mathcal{L}}\rangle_{\mathcal{H}}$ is maximized over $h \in \mathcal{F}$ by the unit-norm function aligned with $g_{\mu,\mathcal{L}}$, that is, by $g_{\mu,\mathcal{L}}/\|g_{\mu,\mathcal{L}}\|_{\mathcal{H}} \in \mathcal{F}$. The squared RKHS norm $\|g_{\mu,\mathcal{L}}\|_{\mathcal{H}}^2$ can be written in terms of the kernel as

$$
\begin{aligned}
\|g_{\mu,\mathcal{L}}\|_{\mathcal{H}}^2 = \langle g_{\mu,\mathcal{L}}, g_{\mu,\mathcal{L}}\rangle_{\mathcal{H}} &= \mathbb{E}_{\mathbf{x}\sim\mu}\big[\mathcal{L}_{\mathbf{x}}g_{\mu,\mathcal{L}}(\mathbf{x})\big] && \text{Lemma 1} \\
&= \mathbb{E}_{\mathbf{x}\sim\mu}\big[\mathcal{L}_{\mathbf{x}}\big[\mathbb{E}_{\mathbf{x}'\sim\mu}\big[\mathcal{L}_{\mathbf{x}'}k(\mathbf{x}',\mathbf{x})\big]\big]\big]\,. && \text{Explicit form of } g_{\mu,\mathcal{L}}(\cdot)
\end{aligned}
$$

We call this expression the *kernel deviation from stationarity* $\mathrm{KDS}(\mathcal{L},\mu;\mathcal{F})$.

Under additional regularity conditions on f, σ, k, and μ, we may interchange the differentials in $\mathcal{L}_{\mathbf{x}}$ with the integral in $\int\mu(\mathbf{x}')\mathcal{L}_{\mathbf{x}'}k(\mathbf{x}',\mathbf{x})\mathrm{d}\mathbf{x}'$ (or specifically, their involved limits) and write $\mathrm{KDS}(\mathcal{L},\mu;\mathcal{F}) = \mathbb{E}_{\mathbf{x}\sim\mu,\mathbf{x}'\sim\mu}\big[\mathcal{L}_{\mathbf{x}}\mathcal{L}_{\mathbf{x}'}k(\mathbf{x},\mathbf{x}')\big]$. For example, by the dominated convergence theorem, one sufficient condition allowing the interchange is when the functions and their first- and second-order partial derivatives are continuous and bounded. More general conditions are possible.

■

B.3 Proof of Theorem 3

Let $\mathcal{H}$ be the RKHS of the Matérn kernel $k_{\nu,\gamma}$, and let $\mathcal{F}$ be its unit ball. This proof uses Lemmata 4 and 5, two auxiliary results about Matérn and Sobolev spaces that are given in Appendix A.

To begin, we note that f, σ, $\partial/\partial x_{i,i} k_{\nu,\gamma}(\mathbf{x},\mathbf{x})$, and $\partial^2/\partial x_{i,i}\partial x_{j,j} k_{\nu,\gamma}(\mathbf{x},\mathbf{x})$ are all square-integrable with respect to μ, because the functions are bounded, and any bounded function is square-integrable with respect to a probability density. Both functions f and σ are bounded by assumption. Moreover, Lemma 5 and $\nu > 2$ imply that the partial derivatives $\partial/\partial x_{i,i} k_{\nu,\gamma}(\mathbf{x},\mathbf{x})$ and $\partial^2/\partial x_{i,i}\partial x_{j,j} k_{\nu,\gamma}(\mathbf{x},\mathbf{x})$ exist and are finite. These functions of $\mathbf{x}$ are bounded, because the Matérn kernel function depends only on the distance between its inputs, which is $\|\mathbf{x}-\mathbf{x}\|_2 = 0$ for any $\mathbf{x}$, and thus these partial derivatives are constant with respect to $\mathbf{x}$. Given the square-integrability of f, σ, $\partial/\partial x_{i,i} k_{\nu,\gamma}(\mathbf{x},\mathbf{x})$, and $\partial^2/\partial x_{i,i}\partial x_{j,j} k_{\nu,\gamma}(\mathbf{x},\mathbf{x})$, all assumptions of Lemma 1 and Theorem 2 are satisfied.

To prove the theorem, we leverage the fact that the smooth functions with compact support C_c^∞ form a core for the generator $\mathcal{A}$ associated to the SDEs when f, σ are Lipschitz continuous and bounded and the matrix $\sigma(\mathbf{x})\sigma(\mathbf{x})^\top$ is positive definite for all $\mathbf{x} \in \mathbb{R}^d$ (Ethier and Kurtz, 1986, Theorem 1.6, p. 370). We can link the core C_c^∞ to the Matérn RKHS $\mathcal{H}$:

Lemma 6 *C_c^∞ is a dense subset of $\mathcal{H}$ with respect to the Sobolev norm $\|\cdot\|_{\nu+d/2,2}$.*

Proof of Lemma 6. The space $W_c^{m,p}$ is defined as the closure of C_c^∞ in the Sobolev space $W^{m,p}$ (Appendix A.2). Therefore, the core C_c^∞ is dense in $W_c^{m,p}$ with respect to the Sobolev norm $\|\cdot\|_{m,p}$. Moreover, $W_c^{m,p} = W^{m,p}$ when both spaces are defined over $\mathbb{R}^d$ (Adams and Fournier, 2003, Corollary 3.23), so C_c^∞ is dense in $W^{m,p}$. From Lemma 4, we know that $W^{\nu+d/2,2} = \mathcal{H}$ for the set of functions. Hence, C_c^∞ is dense in $W^{\nu+d/2,2} = \mathcal{H}$ with respect to the Sobolev norm $\|\cdot\|_{\nu+d/2,2}$.

We now prove both directions of the equivalence in the theorem:

$\Leftarrow$ If $\mathrm{KDS}(\mathcal{L},\mu;\mathcal{F}) = 0$, then $\sup_{h\in\mathcal{F}} \mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}h(\mathbf{x})] = 0$ by Theorem 2. Since the supremum is nonnegative, it follows that $\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}h(\mathbf{x})] = 0$ for all $h \in \mathcal{F}$. This implies that the equality also holds for $h \in \mathcal{H}$, since the length of the vectors does not affect their orthogonality. When $\|h\|_\mathcal{H} > 0$, we can also see this from $\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}h(\mathbf{x})] = \langle h, g_{\mu,\mathcal{L}}\rangle_\mathcal{H} = \|h\|_\mathcal{H}\langle h/\|h\|_\mathcal{H}, g_{\mu,\mathcal{L}}\rangle_\mathcal{H} = \|h\|_\mathcal{H}\cdot 0 = 0$ since $h/\|h\|_\mathcal{H} \in \mathcal{F}$.

By Lemma 6, the core C_c^∞ is a subset of $\mathcal{H}$, so we have $\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}u(\mathbf{x})] = 0$ for all $u \in C_c^\infty$. If $u \in C_c^\infty$, then $u \in C_c^2$ and thus $\mathcal{A}u = \mathcal{L}u$. It follows that $\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{A}u(\mathbf{x})] = 0$ for all u in the core C_c^∞. This implies that μ is the stationary density (Ethier and Kurtz, 1986, Chapter 4, Proposition 9.2).

$\Rightarrow$ If μ is the stationary density, we have $\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{A}u(\mathbf{x})] = 0$ for all functions u in the core C_c^∞. Moreover, since $C_c^\infty \subset C_c^2$, it holds that $\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}u(\mathbf{x})] = 0$.

Let $h \in \mathcal{H}$. By Lemma 6, there exists $u \in C_c^\infty$ such that $\|h-u\|_{\nu+d/2,2} < \epsilon$. By the above, we then have

$$
\begin{aligned}
|\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}h(\mathbf{x})]| &= |\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}h(\mathbf{x}) - \mathcal{L}u(\mathbf{x}) + \mathcal{L}u(\mathbf{x})]| && \text{expanding}\\
&\le |\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}(h-u)(\mathbf{x})]| + |\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}u(\mathbf{x})]| && \text{triangle inequality}\\
&= |\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}(h-u)(\mathbf{x})]|\\
&= |\langle h-u, g_{\mu,\mathcal{L}}\rangle_\mathcal{H}| && \text{Lemma 1}\\
&\le \|h-u\|_\mathcal{H}\,\|g_{\mu,\mathcal{L}}\|_\mathcal{H} && \text{Cauchy-Schwarz}\\
&\le c_2\|h-u\|_{\nu+d/2,2}\,\|g_{\mu,\mathcal{L}}\|_\mathcal{H} && \text{Lemma 4}\\
&< c_2\,\epsilon\,\|g_{\mu,\mathcal{L}}\|_\mathcal{H}
\end{aligned}
$$

Thus, $|\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}h(\mathbf{x})]|$ is bounded by ϵ times the constants c_2 and $\|g_{\mu,\mathcal{L}}\|_\mathcal{H}$, which are both independent of the function h. Hence, for all functions $h \in \mathcal{H}$ and any $\epsilon' > 0$, we can choose $\epsilon > 0$ such that $|\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}h(\mathbf{x})]| < \epsilon'$. It follows that $\mathbb{E}_{\mathbf{x}\sim\mu}[\mathcal{L}h(\mathbf{x})] = 0$ for all $h \in \mathcal{H}$ and, by Theorem 2, $\mathrm{KDS}(\mathcal{L},\mu;\mathcal{F}) = 0$.

■

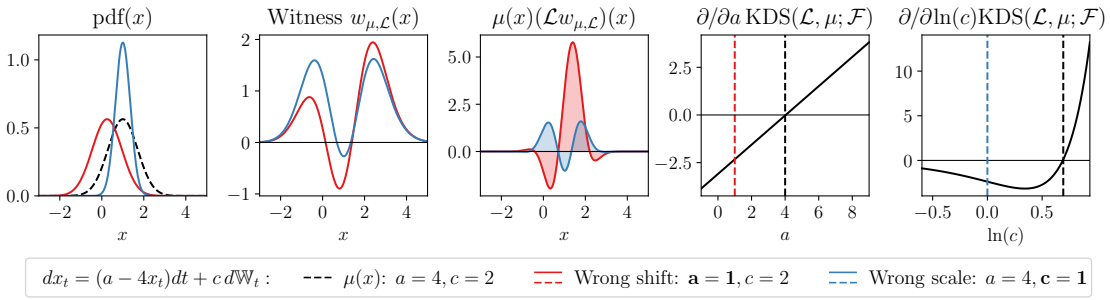


Figure 3: Components of the KDS for a stationary linear SDE and a Gaussian kernel k_γ with $\gamma = 0.5$. Expectations over μ are approximated with 1000 samples. *1:* Densities of a target (μ, black) and two alternative models. *2:* KDS witness functions for the misspecified models. *3:* Witnesses after applying $\mathcal{L}$, yielding their time derivatives in the diffusion. After multiplying by μ, the KDS is equal to the integral of the shaded areas. *4-5:* KDS derivatives with respect to a and c, fixing the other parameters at those of the target model. The partial derivatives have zeroes at the true parameters of the model inducing μ, thus gradient descent drives the incorrect a and c to their true values (indicated by vertical, dashed lines).

C Additional Details on the Kernel Deviation from Stationary

C.1 Explicit Form

It is instructive to consider the special case of $\sigma = \mathbf{I}$. The KDS function $\mathcal{L}^{\boldsymbol{\theta}}_{\mathbf{x}}\mathcal{L}^{\boldsymbol{\theta}}_{\mathbf{x}'}k$ is then given by

$$\begin{aligned}\mathcal{L}^{\boldsymbol{\theta}}_{\mathbf{x}}\mathcal{L}^{\boldsymbol{\theta}}_{\mathbf{x}'}k(\mathbf{x},\mathbf{x}') = f_{\boldsymbol{\theta}}(\mathbf{x})\cdot\nabla_{\mathbf{x}}\nabla_{\mathbf{x}'}k(\mathbf{x},\mathbf{x}')\cdot f_{\boldsymbol{\theta}}(\mathbf{x}') + \tfrac{1}{2}f_{\boldsymbol{\theta}}(\mathbf{x})\cdot\nabla_{\mathbf{x}}\Delta_{\mathbf{x}'}k(\mathbf{x},\mathbf{x}') \\ + \tfrac{1}{2}f_{\boldsymbol{\theta}}(\mathbf{x}')\cdot\nabla_{\mathbf{x}'}\Delta_{\mathbf{x}}k(\mathbf{x},\mathbf{x}') + \tfrac{1}{4}\Delta_{\mathbf{x}}\Delta_{\mathbf{x}'}k(\mathbf{x},\mathbf{x}')\,,\end{aligned} \tag{14}$$

where $\Delta_{\mathbf{x}} := \operatorname{tr}\nabla_{\mathbf{x}}\nabla_{\mathbf{x}}$ is the Laplacian. This expression contains a matrix, two vectors, and a scalar involving k that are all independent of the model parameters $\boldsymbol{\theta}$. Thus, we can precompute and reuse these kernel terms for any $\boldsymbol{\theta}$, e.g., during optimization of $\boldsymbol{\theta}$ with gradient descent. For general $\sigma_{\boldsymbol{\theta}}$, there also exists an explicit expression, but it may be easier to leverage the operator view of $\mathcal{L}^{\boldsymbol{\theta}}$ and compute the gradients of $\mathcal{L}^{\boldsymbol{\theta}}_{\mathbf{x}}\mathcal{L}^{\boldsymbol{\theta}}_{\mathbf{x}'}k$ with automatic differentiation. We provide the explicit form and pseudocode demonstrating this case in Appendix C.

C.2 Example

Figure 3 illustrates how the KDS may be used to learn the SDE parameters $\boldsymbol{\theta}$. We consider an instance of a target linear model $dx_t = (a + bx_t)dt + c\mathrm{d}\mathbb{W}_t$ with the closed-form density $\mu(x) = \mathcal{N}(x; -a/b, -c^2/2b)$ (Jacobsen, 1993) for $b < 0$ and $c > 0$. We use the KDS, approximated by samples from μ, to measure the fit of two models with incorrect a and c controlling the mean and variance, respectively. The partial derivatives of the KDS have zeroes at the true parameters of the model inducing μ and can thus be inferred with gradient descent (details in Figure 3).

D Experimental Setup

D.1 Data

D.1.1 Sparsity Structures

For benchmarking, we simulate data from randomly-generated sparse linear systems and sparse gene regulatory network models with $d = 20$ variables. Following prior work (e.g., Zheng et al., 2018), we sample random sparsity structures $\mathbf{G} \in \{0,1\}^{d\times d}$ with either polynomial or power-law degree distributions of the variables, corresponding to Erdős-Rényi and scale-free graphs, respectively. Erdős-Rényi graphs are sampled by drawing links independently with a fixed probability (when acyclic, restricted to an upper-triangular matrix). Scale-free graphs are generated by a sequential preferential attachment process, where links of node j to the previous $j-1$ nodes are sampled with probability proportional to its degree and then randomly directed (when acyclic, always directed ingoing to j). For both sparsity models, we fix the (expected) degree of the variables to 3.

D.1.2 Cyclic Linear Systems

Models Given some $\mathbf{G} \in \{0,1\}^{d \times d}$, we generate random instances of the two cyclic linear models

$$\mathbf{x} = \mathbf{W}\mathbf{x} + \mathbf{b} + \mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{\epsilon} \;\text{ with } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \qquad \text{(Cyclic Linear SCM)}$$

$$\mathrm{d}\mathbf{x}_t = (\mathbf{W}\mathbf{x}_t + \mathbf{b})\mathrm{d}t + \mathrm{diag}(\boldsymbol{\sigma})\mathrm{d}\mathbb{W}_t \qquad \text{(Cyclic Linear SDE)}$$

where $\mathbf{W} \in \mathbb{R}^{d \times d}$, $\mathbf{b} \in \mathbb{R}^d$, and $\boldsymbol{\sigma} \in \mathbb{R}^d_{>0}$, and $\mathbf{W}$ is sparse according to $\mathbf{G}$. Sampling random cyclic systems requires more caution than in the acyclic case, since both generative processes must be stable. For SCMs, the maximum of the real-parts of the eigenvalues $\rho(\mathbf{W})$ must be less than 1, for SDEs less than 0. For an insightful evaluation, we additionally want $\mathbf{W}$ to be asymmetric and not approximately diagonal, i.e., have significant causal dependencies between the variables.

To generate such systems, we first sample $\mathbf{G} \sim p(\mathbf{G})$, $\mathbf{W} \sim p(\mathbf{W})$, $\mathbf{b} \sim p(\mathbf{b})$, $\boldsymbol{\sigma} \sim p(\boldsymbol{\sigma})$. Then, we multiply $\mathbf{W}$ times $\mathbf{G}$ elementwise along their offdiagonal elements and finally subtract $\rho(\mathbf{W}) + \epsilon$ from the diagonal of $\mathbf{W}$, which ensures that $\rho(\mathbf{W}) \leq -\epsilon$. We found that matrices $\mathbf{W}$ sampled by this protocol empirically induce stronger variable correlations in the stationary distributions than the procedure by Varando and Hansen (2020). They perform a more vacuous diagonal shift based on the Gershgorin circle theorem, often resulting in large dominating diagonals. For our experiments, we use $p(w_{ij}) = \mathrm{Unif}(-3,-1) \cup (1,3)$ and $\epsilon = 0.5$ for the matrices and $p(b_j) = \mathrm{Unif}(-3,3)$ and $p(\log \sigma_j) = \mathrm{Unif}(-1,1)$ for the biases and scales, both for the SCMs and SDEs.

To sample the SCM data, we draw $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and then compute $\mathbf{x} = (\mathbf{I} - \mathbf{W})^{-1}(\mathbf{b} + \mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{\epsilon})$ (Hyttinen et al., 2012). To sample from the stationary density of the SDEs, we use the Euler-Maruyama scheme (Appendix A.1).

Interventions Given the fully-specified linear model, we sample an observational dataset and interventional data for single-variable shift interventions on all variables, each with 1000 observations, as the interventions for the benchmark. In both SCMs and SDEs, the shift intervention is implemented by adding a scalar δ to the bias b_j of the target variable j. In our experiments, we sample $\delta \sim p(\delta)$ with $p(\delta) = \mathrm{Unif}(-15,-5) \cup (5,15)$ independently for each intervention.

D.1.3 Gene Regulatory Networks

Model Given some acyclic $\mathbf{G} \in \{0,1\}^{d \times d}$, we use the SERGIO model by Dibaeinia and Sinha (2020) and their corresponding implementation (GNU General Public License v3.0) to sample synthetic gene expression data. The gene expressions are simulated by a stationary dynamical system over a sparse, acyclic regulatory network encoded by $\mathbf{G}$. To simplify the experimental setup, we use the clean gene expressions without technical measurement noise as the observations.

SERGIO models the mRNA concentration of the genes using the chemical Langevin equation, a nonlinear geometric Brownian motion model driven by two independent Wiener processes for each gene. The expression x_j of gene j is primarily defined through its production rate p_j, which depends nonlinearly on the expression levels $\mathbf{x}$ of the other genes through the signed interaction parameters $\mathbf{K}$ and the regulatory network $\mathbf{G}$. Following Dibaeinia and Sinha (2020), we use a Hill nonlinearity coefficient of 2 and sample the parameters k_{ij} as well as 10 master regulator rates b_{jc}, which model cell type heterogeneity, from $k_{ij} \sim \mathrm{Unif}(-5,-1) \cup (1,5)$ and $b_{jc} \sim \mathrm{Unif}(1,4)$, respectively. Finally, we use an expression decay rate of $\lambda = 0.5$ and noise scale of $q = 0.5$, which deviates from the values 0.8 and 1.0, respectively, used by Dibaeinia and Sinha (2020) when simulating $d \geq 100$ genes. Under their settings, the data of smaller networks does not contain sufficient signal for the any of the benchmarked methods to learn a nontrivial model of the system.

Interventions Given the fully-specified gene regulation model, we sample an observational (wild-type) dataset and interventional data for single-variable gain-of-function (overexpression) interventions (e.g., Norman et al., 2019) on all genes, each with 1000 measured cell observations, as the interventions for the benchmark. We evaluate overexpression rather than knockdown perturbations, because the former are qualitatively more similar to the test-time shift interventions used to query the models learned by the methods. The gain-of-function interventions are implemented by multiplying the production rate p_j of the target gene j by a randomly-sampled factor $r_j \sim \mathrm{Unif}(2,10)$. The half-response levels for the Hill nonlinearities are kept at the values estimated during the wild-type simulation, so that the intervention effects propagate downstream.

D.2 Metrics

We focus on comparing the true and predicted interventional distributions of unseen interventions in a system. While this evaluation setting mimics real-world applications, benchmarking different algorithms requires some care. In general, there is a mismatch between the model-level perturbation implemented by an intervention in the ground-truth system and the *query* perturbation performed in a learned causal model—not only because the true model perturbation is unknown, but also because true and learned models may be from different model classes.

Test-time interventions To compare algorithms at test-time, we perturb each learned model by a shift intervention on the target variable that induces the same target variable mean as the true, held-out perturbation data (Rothenhäusler et al., 2015; Zhang et al., 2021). We perform shift interventions because they have analogous implementations in both additive-noise SCMs (1) and stationary diffusions (10) by adding a scalar δ to the mechanism $f_j(\mathbf{x})$ of the target variable j (see also Appendix D.1.2). After performing the intervention, our metrics compare the predicted and true interventional joint distributions. To make this protocol well-defined, we assume knowledge of the true interventional mean of the target variable.[2]

For acyclic SCMs with additive noise, the test-time shift δ required for the query perturbation is directly given by the difference between the empirical observational mean of the learned SCM and the target interventional mean. However, cyclic SCMs and stationary diffusions may model feedback on the target variable, where the above does not hold. For cyclic models, we individually find the query shift δ by performing an exponential search around $\delta = 0$ for a range estimate $(\delta_{\text{lo}}, \delta_{\text{hi}})$. At each shift value, we simulate data from the perturbed model and compare the predicted to the target interventional mean. Given an estimated range $(\delta_{\text{lo}}, \delta_{\text{hi}})$, we run a final grid search for the optimal value $\delta \in \{\delta_{\text{lo}}, \delta_{\text{lo}} + 1/10(\delta_{\text{hi}} - \delta_{\text{lo}}), \dots, \delta_{\text{lo}} + 9/10(\delta_{\text{hi}} - \delta_{\text{lo}}), \delta_{\text{hi}}\}$. Ultimately, we select the shift δ achieving the $\arg\min$ distance to the target interventional mean in the grid search.

Metrics Both metrics we report are computed based on *samples* from the predicted distributions, which enables a nonparametric comparison across the different probabilistic models. For each test intervention, we simulate 1000 samples $\hat{\mathbf{x}}^{(i)} \in \hat{D}$ from the interventional distribution of the predicted model and compare them with the true interventional dataset of 1000 samples $\mathbf{x}^{(i)} \in D$.

To evaluate the overall fit of the predicted data distribution, we compute the Wasserstein distance W_2 to the ground-truth interventional data. To make W_2 efficiently computable, we report the W_2 distance with small entropic regularization, which interpolates between the W_2 distance and the MMD (Genevay et al., 2018) and commonly serves as an evaluation metric in machine learning applications (e.g., Bunne et al., 2022). The entropy-regularized W_2 metric between the empirical measures of the datasets $\hat{D}$ and D with $|\hat{D}| = M$ and $|D| = N$ is defined as

$$W_2(\hat{D}, D) := \left(\min_{\mathbf{P} \in U} \sum_{m=1}^{M} \sum_{n=1}^{N} p_{mn} \|\hat{\mathbf{x}}^{(m)} - \mathbf{x}^{(n)}\|_2^2 - \epsilon H[\mathbf{P}] \right)^{1/2},$$

where H is the entropy defined as $H[\mathbf{P}] := -\sum_{nm} p_{mn}(\log p_{nm} - 1)$, and U is the set of transport matrices $U = \{\mathbf{P} \in \mathbb{R}_{\geq 0}^{M \times N} : \mathbf{P}\mathbf{1}_N = 1/M\, \mathbf{1}_M \text{ and } \mathbf{P}^\top \mathbf{1}_M = 1/N\, \mathbf{1}_N\}$ with $\mathbf{1}_N$ being a vector of N ones (Peyré et al., 2019). We found the W_2 metric to be a more robust evaluation metric than the MMD, because it does not depend on the sensitive choice of a kernel bandwidth (Gretton et al., 2012). For $\epsilon > 0$, the entropy-regularized W_2 distance can be efficiently computed using the Sinkhorn algorithm, which we use as implemented by the `ott-jax` package (Apache 2.0 Licence) with $\epsilon = 0.1$ (Cuturi et al., 2022).

To separately assess the accuracy of the interventional means, we follow Zhang et al. (2022) and report the mean squared error of the predicted empirical means of the d variables given by

$$\text{MSE}(\hat{D}, D) := \frac{1}{d} \sum_{j=1}^{d} (\hat{m}_j - m_j)^2,$$

where $\hat{\mathbf{m}} := \frac{1}{M} \sum_{m=1}^{M} \hat{\mathbf{x}}^{(m)}$ and $\mathbf{m} := \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}^{(n)}$ are the empirical means of the datasets.

[2] Outside benchmarking settings, in which we compare to a ground-truth reference dataset, the 'true' intervention effect on the target is always known, because it corresponds to the query we pose to the learned causal model (e.g., when asking: "what is the genome-wide effect of over-expressing gene x_j two-fold?")

D.3 Hyperparameter Tuning

In the experiments, we benchmark the methods on different generative processes (Appendix D.1). To calibrate the important hyperparameters of the methods, we perform cross-validation prior to the final evaluation that benchmarks the methods. All methods are tuned separately for each data-generating process, that is, for cyclic linear SCMs, cyclic linear SDEs, and the synthetic gene expression data, both for Erdős-Rényi and scale-free sparsity structures.

The experiments for all generative processes are repeated for 50 randomly-sampled systems (Figure 2, Section 6 and Figure 4, Appendix E). Each task instance consists of an observational and 10 interventional datasets for learning the model as well as 10 interventional datasets for the final evaluation, with all interventions performed on separate target variables. To tune the hyperparameters of the methods, we split the 10 observed interventions into 9 training and 1 validation dataset. The methods then infer a causal model based on the 9 training interventional and the observational dataset, and we compute the W_2 metric for the unseen validation intervention. For each method, we select the hyperparameter configuration achieving the lowest median W_2 metric on 20 randomly-selected tasks.

D.4 Stationary Diffusions

Models We evaluate linear and nonlinear stationary diffusion models. Both classes of SDE systems model d independent drift and diffusion mechanisms f_j and σ_j that are defined by separate parameters $\boldsymbol{\theta}_j$, as in (8). For both models, the corresponding group lasso regularizers $R(\boldsymbol{\theta}_j)$ penalize the dependence on the other variables. The models and regularizers are defined as

$$
\begin{aligned}
f_{\boldsymbol{\theta}_j}(\mathbf{x})_j &= b^j + \mathbf{w}^j \cdot \mathbf{x} & R(\boldsymbol{\theta}_j) &= \textstyle\sum_{i \neq j}^{d} |w_i^j| \\
f_{\boldsymbol{\theta}_j}(\mathbf{x})_j &= b^j + \mathbf{w}^j \cdot g(\mathbf{U}^j \mathbf{x} + \mathbf{v}^j) - x_j & R(\boldsymbol{\theta}_j) &= \textstyle\sum_{i \neq j}^{d} \|\mathbf{u}_i^j\|_2
\end{aligned}
$$

where $g(z) := \exp(z)/(\exp(z)+1)$ the sigmoid nonlinearity, applied elementwise. The diffusion term σ is modeled as a constant matrix with $\sigma(\mathbf{x}) = \operatorname{diag}(\exp(\log \boldsymbol{\sigma}))$, $\log \boldsymbol{\sigma} \in \mathbb{R}^d$. The parameters $\log \boldsymbol{\sigma}$ are learned in log-space to enable gradient-based optimization while respecting $\sigma(\mathbf{x}) \in \mathbb{R}^d_{>0}$. To remove the speed scaling invariance, we fix $w_j^j = -1$ in the linear and $\mathbf{u}_j^j = \mathbf{0}$ in the MLP model (see Section 4.4). In the experiments, the MLP model uses a hidden size of $h = 8$ for the matrices $\mathbf{U}^j \in \mathbb{R}^{h \times d}$ and vectors $\mathbf{v}^j, \mathbf{w}^j \in \mathbb{R}^h$.

Interventions during training In all evaluation settings and for both diffusion models, we jointly learn shift interventions $\boldsymbol{\phi}_j = \{\delta_j\}$ as defined in (10, left) for the target variables of each training environment (see Section 4.4). For the purpose of the experiments, we limit the interventions to shifts in order to allow a direct comparison with SCMs. However, we found learning more complex intervention parameterizations like, for example, full shift-scale interventions as in (10), generally straightforward. More expressive interventions shift some of the burden of explaining the distribution shift from $\boldsymbol{\theta}$ to $\boldsymbol{\phi}_j$, which can help inferring robust parameters $\boldsymbol{\theta}$ under model mismatch.

Optimization We learn the model $\boldsymbol{\theta}$ and the intervention parameters $\{\boldsymbol{\phi}_j\}$ jointly with gradient-based optimization. By default for all experiments, we run 20,000 update steps on the KDS as described in Section 4.4 using the Adam optimizer with learning rate 0.001. We compute the empirical KDS (6) using the Gaussian kernel k_γ and a batch size of $|D| = 512$. The parameters $\boldsymbol{\theta}$ are initialized near zero by sampling from $\mathcal{N}(0, 0.001^2)$. We warm-start the intervention shifts $\boldsymbol{\phi}_j = \{\delta_j\}$ by initializing them at the difference in means of the target variable in the interventional and the observational datasets. Overall, the important hyperparameters are the kernel bandwidth γ and the group lasso regularization strength λ, so we tune these for each experimental setting via a grid search (see Table 1) using the protocol described in Appendix D.3.

Diagnostics The following intuitions may be helpful when deploying our inference approach. If the stationary density induced by the learned SDEs overfits or collapses to a small part of the data distribution, then the kernel bandwidth may be too small. The bandwidth range searched over in our experiments is suitable for standardized datasets of $d = 20$ variables but should likely be expanded in different settings. If the learned SDEs are unstable upon convergence or diverge during test simulations late in training—despite a decreasing or near-zero KDS loss—then the speed scaling invariance may not be adequately fixed (see above and Section 4.4). In this context, we find that the

Table 1: Hyperparameter tuning for the experiments in Section 6. The hyperparameters of all methods are selected using the protocol described in Appendix D.3.

Method	Hyperparameter	Range
IGSP	significance level	$\alpha_{\text{IGSP}} \in \{0.001, 0.003, 0.01, 0.03, 0.1\}$
DCDI	sparsity regularization	$\lambda_{\text{DCDI}} \in \{0.001, 0.01, 0.1, 1, 10\}$
	number of MLP layers	$m_{\text{DCDI}} \in \{1, 2\}$
NODAGS	sparsity regularization	$\lambda_{\text{NODAGS}} \in \{0.0001, 0.001, 0.01, 0.1\}$
	spectral norm terms	$n_{\text{NODAGS}} \in \{5, 10, 15\}$
	learning rate	$\eta_{\text{NODAGS}} \in \{0.001, 0.01, 0.1\}$
	hidden units	$m_{\text{NODAGS}} \in \{1, 2, 3\}$
LLC	sparsity regularization	$\lambda_{\text{LLC}} \in \{0.001, 0.01, 0.1, 1, 10, 100\}$
KDS	sparsity regularization	$\lambda \in \{0.001, 0.003, 0.01, 0.03, 0.1\}$
	kernel bandwidth	$\gamma \in \{3, 5, 7\}$

fit and performance of the models empirically improves when fixing the self-regulating parameters of f_j on x_j, rather than, e.g., the noise scales σ_j. Without any sparsity regularization, gradient descent as in (11) may converge to models at the edge of stability, e.g., to linear models with maximum real parts of the eigenvalues being near zero and only just negative. Sparsity regularization can mitigate such instability and related issues in combination with fixing the speed scaling.

D.5 Baselines

GIES (Hauser and Bühlmann, 2012) assumes a linear-Gaussian SCM to infer a graph equivalence class, from which we randomly sample a causal graph. To perform the greedy search, we run the original `R` implementation of the authors using the Causal Discovery Toolbox (MIT Licence)[3]. Given the DAG estimate, we use a linear-Gaussian SCM with maximum likelihood parameter and variance estimates as the learned model. These estimates have simple closed-forms that account for interventional data (Hauser and Bühlmann, 2012). At test time, the shift interventions are implemented in the learned linear SCM and the data sampled as described in Appendix D.1.2.

IGSP (Wang et al., 2017) uses a Gaussian partial correlation test. We use the same closed-form maximum likelihood parameter and variances estimates as for GIES to construct the final causal model. For IGSP, we run the implementation provided as part of the CausalDAG package (3-Clause BSD License)[4]. Using the protocol described in Appendix D.3, we tune the significance level α_{IGSP} of the conditional independence test for each experimental setting individually by searching over a range of α_{IGSP} values (see Table 1). As for GIES, the shift interventions are implemented in the estimated linear SCM as described in Appendix D.1.2.

DCDI (Brouillard et al., 2020) learns a nonlinear, Gaussian SCM parameterized by neural networks jointly with the noise variance. For comparison with the nonlinear stationary diffusion model, we use the same hidden size of 8 for the neural networks. To run DCDI, we use the Python implementations provided by the authors (MIT License). We tune the regularization strength λ_{DCDI} and the number of layers m_{DCDI} and otherwise leave the remaining optimization hyperparameters at the suggestions by the authors (see Table 1). When learning from imperfect interventions, DCDI estimates a separate model for each interventional environment. To evaluate the performance on unseen interventions, we use the model learned for the observational dataset and implement the shift interventions by adding the bias δ to the mean of the Gaussian modeling the target variable, analogous to the linear SCMs and SDEs described in Appendix D.1.2.

NODAGS (Sethuraman et al., 2023) infers a nonlinear cyclic SCM using residual normalizing flows and also estimates the noise variances. As suggested by the authors, we jointly tune the

[3] `https://github.com/FenTechSolutions/CausalDiscoveryToolbox`
[4] `https://github.com/uhlerlab/causaldag`

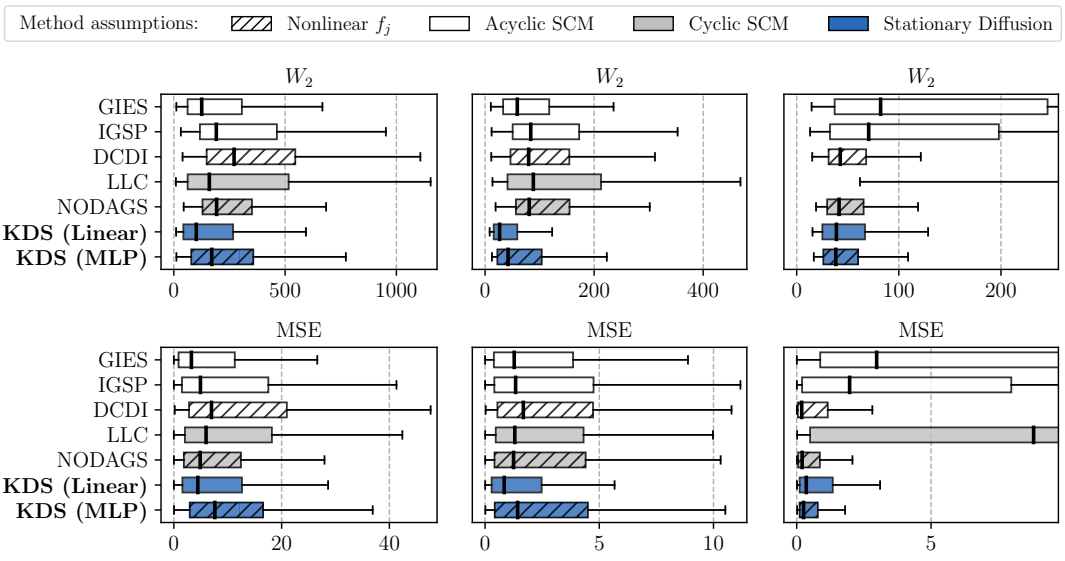
regularization parameter λ_{NODAGS}, the number of terms for computing the spectral norm n_{NODAGS}, the learning rate η_{NODAGS}, and the number of hidden units m_{NODAGS} (see Table 1). We set the remaining hyperparameters to the recommendations of the authors and use the implementation published alongside the original paper (Apache 2.0 Licence). At test time, the shift interventions are implemented in the model by using $\mathbf{U} = \mathbf{I}$ and otherwise as described in the paper, analogous to the linear cyclic SCMs described in Appendix D.1.2 (see also Hyttinen et al., 2012).

LLC (Hyttinen et al., 2012) learns a linear cyclic SCM and estimates the noise variances. For the basic implementation of the LLC algorithm, we use the code provided by the NODAGS repository. However, we extend their implementation by the ℓ_1 sparsity regularizer described in Section 6.2 of the original paper by Hyttinen et al. (2012), solving the minimization problem with BFGS. We treat the weight λ_{LLC} of this regularizer as a hyperparameter that is tuned via a grid search (see Table 1). At evaluation time, the shift interventions in the learned cyclic linear SCM are performed as for GIES and IGSP.

D.6 Compute Infrastructure

The development and experiments of this work were carried out on an internal cluster. In each experiment, all methods ran for up to 1 hour of wall time on up to 4 CPUs and 16 GB of RAM, adjusted individually according to the compute requirements of each method. We implement our approach with `JAX` (Bradbury et al., 2018) and thus additionally provide 1 GPU, which allows for significant speed-ups during development and the experiments. Overall, running our inference method takes approximately one hour given the above resources, both for the linear and nonlinear model, and including the final search for test-time intervention shifts.

E Additional Results



(a) **Cyclic Linear SCMs** (b) **Cyclic Linear SDEs** (c) **SERGIO**

Figure 4: Benchmarking results ($d = 20$ variables, scale-free sparsity structure). Metrics are computed from 10 test interventions on unseen target variables in 50 randomly-generated systems. Box plots show medians and interquartile ranges (IQR). Whiskers extend to the largest value inside 1.5 times the IQR length from the boxes.