Journal of Machine Learning Research 27 (2026) 1-56 Submitted 3/25; Revised 8/26; Published 9/26

Identifying Drift, Diffusion, and Causal Structure from Temporal Snapshots

Vincent Guan[1*] VGUAN23@MATH.UBC.CA
Joseph Janssen[2*] JOEJANSSEN@EOAS.UBC.CA
Hossein Rahmani[1] HRAHMANI@MATH.UBC.CA
Andrew Warren[1] AWARREN@MATH.UBC.CA
Stephen Y. Zhang[3] SYZ@SYZ.ID.AU
Elina Robeva[1] EROBEVA@MATH.UBC.CA
Geoffrey Schiebinger[1] GEOFF@MATH.UBC.CA

[1] *Department of Mathematics, University of British Columbia*
[2] *Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia*
[3] *School of Mathematics and Statistics, University of Melbourne*
[*] *Corresponding author*

Editor: Bryon Aragam

Abstract

Stochastic differential equations (SDEs) are a fundamental tool for modelling dynamic processes, including gene regulatory networks (GRNs), contaminant transport, financial markets, and image generation. However, learning the underlying SDE from data is a challenging task, especially if individual trajectories are not observable. Motivated by burgeoning research in single-cell datasets, we present the first comprehensive approach for jointly identifying the drift and diffusion of an SDE from its temporal marginals. Assuming linear drift and additive diffusion, we show that non-identifiability can only arise if the initial distribution possesses generalized rotational symmetries. We further prove that even if this condition holds, the drift and diffusion can almost always be recovered from the marginals. Additionally, we show that the causal graph of any SDE with additive diffusion can be recovered from the identified SDE parameters. To complement this theory, we adapt entropy-regularized optimal transport to handle anisotropic diffusion, and introduce APPEX (Alternating Projection Parameter Estimation from X_0), an iterative algorithm designed to estimate the drift, diffusion, and causal graph of an additive noise SDE, solely from temporal marginals, and demonstrate its effectiveness on simulated data from linear additive noise SDEs.

Keywords: stochastic differential equations (SDEs), optimal transport, trajectory inference, causal inference, Schrodinger bridge

1. Introduction

This work presents the first comprehensive approach for jointly identifying the drift and diffusion of a stochastic differential equation (SDE) from observed temporal marginals. While parameter estimation has been studied extensively using trajectory data, either from one long trajectory (Nielsen et al., 2000; Bishwal, 2007), or multiple short trajectories

©2026 Vincent Guan, Joseph Janssen, Hossein Rahmani, Andrew Warren, Stephen Zhang, Elina Robeva and Geoffrey Schiebinger.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-0541.html.

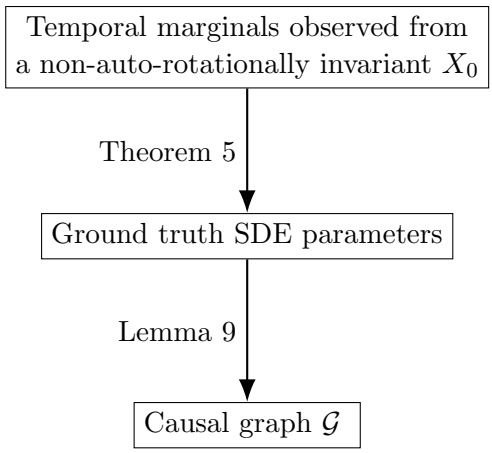
(Manten et al., 2024; Lu et al., 2021; Pawlowicz et al., 2019), it is often impossible to observe individual trajectories when tracking large populations. For example, in single-cell RNA sequencing (scRNA-seq) datasets, the destructive nature of sequencing technologies prevents specific cells from being tracked over time (Trapnell et al., 2014; Setty et al., 2016; Farrell et al., 2018). Similarly, hydrogeochemical sensors cannot track distinct particles when studying contaminant flow, providing only data on plume migration (Man and Tsai, 2007; Hazas et al., 2023; Salamon et al., 2007; Elfeki, 2006; Adams and Gelhar, 1992; Boggs et al., 1992; Mackay et al., 1986; Nenna et al., 2011). Observational data may therefore be limited to samples from temporal marginals of the process at various measurement times. We refer to this observational setting as *marginals-only*.

Many stochastic processes are not identifiable under the marginals-only setting, since their distributions may evolve identically, despite having distinct parameter sets (Weinreb et al., 2018). Previous works therefore simplify the problem by assuming that the drift is irrotational and that the diffusion is isotropic, with known diffusivity (Weinreb et al., 2018; Lavenant et al., 2021; Vargas et al., 2021; Chizat et al., 2022).

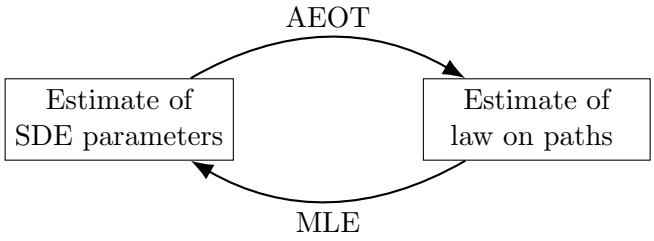
1.1 Our Contributions

We prove that the drift, diffusion, and underlying causal graph of an SDE with linear drift and additive diffusion are generically identifiable (meaning they can be uniquely recovered from the data for almost every SDE parameter configuration) from its temporal marginals. We also introduce a practical algorithm for estimating each of these quantities solely from the marginals of a general additive diffusion process, and demonstrate the effectiveness of our method on simulated data. We illustrate our theoretical contributions in Figure 1a and our algorithmic contributions in Figure 1b. A detailed overview of our contributions is also given below.

1. We provide a theoretical foundation for system identification from temporal marginals.
   (a) In Section 3, we prove identifiability conditions for jointly recovering the drift and diffusion of a linear additive noise SDE from continuously observed marginals. In Theorem 5, we prove that unidentifiable parameter configurations exist if and only if the initial distribution possesses generalized rotational symmetries, a condition we call auto-rotational invariance (ARI). Then, Theorem 6 provides a stronger unconditional guarantee: even when the initial distribution is ARI, the set of unidentifiable configurations has measure zero. Thus, linear additive noise SDEs are generically identifiable.
   (b) In Section 4, we connect parameter estimation to causal structure learning via dynamic structural causal models (Boeken and Mooij, 2024). We show that the causal graph of any additive noise model can be recovered from the identified SDE parameters.
2. In Section 5, we introduce our algorithm: Alternating Projection Parameter Estimation from X_0 (APPEX). It is the first Schrödinger Bridge based method designed to estimate drift and diffusion from temporal marginals of additive noise SDEs, without prior knowledge.



(a) Theoretical contributions: identifying drift, diffusion, and causal structure from temporal marginals



(b) Algorithmic contributions: APPEX algorithm

Figure 1: Outline of our theoretical (a) and algorithmic (b) contributions. In (b), our algorithm Alternating Projection Parameter Estimation from X_0 alternates between Anisotropic Entropy-regularized Optimal Transport (AEOT) and Maximum Likelihood Estimation (MLE).

(a) APPEX alternates between a trajectory inference step, which is a Schrödinger bridge problem, and a parameter estimation step, which is solved via maximum likelihood estimation. With each iteration, APPEX decreases the KL divergence between its inferred trajectories and the law of its estimated SDE.

(b) To solve the Schrödinger Bridge problem for trajectory inference, we adapt entropic optimal transport for parameter dependent transport costs, including arbitrary anisotropic diffusion.

(c) We test APPEX's efficacy across a wide range of experiments in Section 6. We first use APPEX to empirically demonstrate that classical non-identifiable linear additive noise SDEs are resolved when X_0 is initialized appropriately according to Theorem 5. We then demonstrate APPEX's effectiveness over a large-scale experiment on higher dimensional SDEs with randomly generated parameter sets. Our results demonstrate that APPEX identifies arbitrary linear additive noise SDEs with significantly higher accuracy than the widely-used Waddington-OT (WOT) method (Schiebinger et al., 2019). Our final experiments demonstrate that APPEX can be used to identify the system's causal graph, even in the presence of correlated diffusion from latent confounders.

2. Background and Related Work

In this section, we first give an overview of SDEs and a discussion of some real-world applications. We then review relevant literature and techniques for statistical inference

from the "marginals-only" observational setting. Finally, we discuss the problem of non-identifiability of SDEs under this setting.

2.1 Stochastic Differential Equations (SDEs) and Modeling Real-World Systems

The mathematical framework of SDEs, or drift-diffusion processes, originates from statistical mechanics. In such applications, the drift models the velocity of particles flowing in a carrier fluid, while diffusion is driven by Brownian motion, and governs their dispersion due to random collisions (Einstein, 1956; MacInnes and Bracco, 1992). Formally, a d-dimensional time-homogeneous SDE is driven by an m-dimensional Brownian motion W_t, and is parameterized by a drift function $b(x) : \mathbb{R}^d \to \mathbb{R}^d$ and a diffusion function $\sigma(x) : \mathbb{R}^d \to \mathbb{R}^{d\times m}$, such that

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t,\ X_0 \sim p_0, \tag{1}$$

where $b(x)$ and $\sigma(x)$ are Lipschitz to ensure the solution's existence and uniqueness (Fang and Giles, 2020).

By capturing deterministic and stochastic dynamics, SDEs provide a powerful framework for modeling complex systems across a variety of fields. For example, SDEs have been used to model cell differentiation in single-cell biology, where a stem cell changes into another cell type, such as a blood cell or bone cell. The drift can be related to the set of genes that directly regulate the expression of a gene of interest (Zhang, 2024; Aalto et al., 2020; Atanackovic et al., 2024; Zhao et al., 2024; Tejada-Lapuerta et al., 2023), and therefore provides insights into genetic diseases and gene therapies. Meanwhile, the diffusion informs the extent to which cell fates are determined by initial conditions (Forrow, 2024). In hydrological systems, drift is linked to important and unknown properties of the subsurface, such as hydraulic conductivity through average flow velocity (Hazas et al., 2023; Beven et al., 1993), while the diffusion usually describes material heterogeneity or turbulence (Beven et al., 1993), and therefore informs the applicability of popular models, such as Darcy's law (Man and Tsai, 2007; Oh and Tsai, 2010; Lichtner et al., 2002). Together, these parameters determine pollutant fates in hydrological systems, a key concern in safeguarding drinking water sources (O'Connor et al., 2002; Frind et al., 2006; Locatelli et al., 2019; Chen et al., 1999; Paulson, 1997; Cahill et al., 2019). In addition to applications to physical sciences, SDEs have also been prominently used for pricing stocks and options in financial markets (Black and Scholes, 1973) and for image generation in computer vision (Meng et al., 2021).

2.2 Linear Additive Noise SDEs

In this paper, we focus on time-homogeneous additive noise SDEs. These SDEs are characterized by an arbitrary drift function $b(X_t)$ and a possibly anisotropic space-independent diffusion $G \in \mathbb{R}^{d\times m}$, such that

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + G\,\mathrm{d}W_t,\ X_0 \sim p_0. \tag{2}$$

We also refer to $H = GG^\top \in \mathbb{R}^{d\times d}$ as the (observational) diffusion, since it is only possible to identify H (Wang et al., 2024; Pavliotis, 2014). Time-homogeneous linear additive noise

SDEs are of particular interest (du Buisson and Touchette, 2023; Wang et al., 2024; Zhang, 2024), as they generalize the popular Ornstein-Uhlenbeck process. A time-homogeneous linear additive noise SDE obeys the form:

$$\mathrm{d}X_t = AX_t\,\mathrm{d}t + G\,\mathrm{d}W_t, \quad X_0 \sim p_0, \quad A \in \mathbb{R}^{d\times d}. \tag{3}$$

We overview important properties of this SDE in Section A of the appendix, including its closed form solution and transition density.

2.3 Inference from Temporal Marginals

The marginals-only observational setting dates back to at least Aristotle, who dissected bird eggs at various developmental stages, in order to infer the processes and mediators (e.g., heat) behind embryogenesis (Waddington, 1935; Ogle et al., 1882; Aristotle, 1942). If the underlying system is modeled by an SDE, then the evolution of its temporal marginals p_t is defined by its Fokker-Planck equation. For example, a straightforward computation shows that the marginals of a linear additive noise SDE evolve according to the following Fokker-Planck equation:

$$\frac{\partial}{\partial t}p(x,t) = -\nabla\cdot((Ax)p(x,t)) + \frac{1}{2}\nabla\cdot(H\nabla p(x,t)), \quad p(x,0) = p_0(x). \tag{4}$$

Contemporary research in statistical inference from temporal marginals largely comes from single-cell biology. Many works focus on trajectory inference for cell fates (Schiebinger et al., 2019; Lavenant et al., 2021; Yachimura et al., 2023; Chizat et al., 2022), while others infer the causal graph (GRN) (Aalto et al., 2020; Brouillard et al., 2020; Tejada-Lapuerta et al., 2023; Rohbeck et al., 2024), or perform parameter estimation (Chakraborty et al., 2009; Shen et al., 2024). Recent works jointly estimate a subset of these quantities by leveraging their intrinsic relationships. For instance, Vargas et al. (2021) and Shen et al. (2025) are primarily interested in trajectory inference, and iteratively estimate drift and trajectories. Meanwhile, Zhang (2024) iteratively estimates drift and trajectories, while additionally applying permanent interventions on the drift dynamics for network inference.

2.4 Schrödinger Bridge Problem

Given a set of temporal marginals, trajectory inference is the task of estimating a law on paths that obeys each marginal constraint while closely approximating desired reference dynamics. This is commonly implemented by solving the Schrödinger Bridge (SB) problem (Chizat et al., 2022; Shen et al., 2024), defined below for a pair of marginals. Given the transition law $\mathcal{K}$ of a reference process, an initial distribution μ, and a final distribution ν, the SB solution is the law of the stochastic process, which satisfies both marginal constraints, while minimizing relative entropy to $\mathcal{K}$:

$$\pi^* = \mathrm{Proj}^{KL}_{\Pi(\mu,\nu)}(\mathcal{K}) = \operatorname*{arg\,min}_{\pi\in\Pi(\mu,\nu)} D_{KL}(\pi\|\mathcal{K}). \tag{5}$$

The SB problem has also gained interest in the machine learning community, due to its connection to generative diffusion models. Score estimation of time-reversed SDEs is an instance of the SB problem (De Bortoli et al., 2021), and SB solvers are now widely used in generative modeling due to their flexibility (De Bortoli et al., 2021; Chen et al., 2021; Shi et al., 2024; Huang, 2024).

2.5 Identifiability of SDE Parameters

The problem of non-identifiability occurs if there exists an alternative drift-diffusion pair, e.g. $(\tilde{A}, \tilde{H}) \neq (A, H)$ for linear-additive noise SDEs, which shares the same marginals. Equivalently, the same curve of marginals solves both Fokker-Planck equations across all observed times, given initial distribution p_0. Non-identifiability can occur due to many factors. For example, distinct SDEs can share the same stationary distribution, and would hence be non-identifiable from marginals if p_0 is initialized at equilibrium (Lavenant et al., 2021; Shen et al., 2024). Also, if the initial distribution is standard normal $p_0 \sim \mathcal{N}(0, I_d)$, then it would be impossible to distinguish between a pure Brownian motion, and an SDE with an additional rotational vector field (Shen et al., 2024). We overview three classical examples of non-identifiability in Section B of the Appendix.

Because of non-identifiability, previous works require assumptions about the underlying SDE to perform inference. The standard assumptions are that the drift is an irrotational vector field $\nabla\psi$ (Weinreb et al., 2018; Lavenant et al., 2021; Vargas et al., 2021; Chizat et al., 2022; Terpin et al., 2024) and that the diffusion is both known and given by isotropic Brownian motion $\sigma \mathrm{d}W_t$ (Weinreb et al., 2018; Lavenant et al., 2021; Vargas et al., 2021; Chizat et al., 2022; Shen et al., 2025; Zhang, 2024; Terpin et al., 2024). It is also common to leverage additional perturbational data from interventions, and to impose regularization, to better learn the system (Rohbeck et al., 2024; Zhang, 2024; Maddu et al., 2024). However, these constraints impose unrealistic conditions on the data. An irrotational drift field cannot model negative feedback loops or repressilator dynamics, which are common in biological data (Weinreb et al., 2018). Similarly, isotropic diffusion cannot model variable noise scales or correlated noise structures, which commonly arise in real data (Mogensen and Hansen, 2022; Rohbeck et al., 2024; Santos et al., 2024). While most works use a known diffusion estimate to infer drift, Forrow (2024) infers diffusion, but requires a drift estimate. In both cases, the accuracy of the prior estimate is crucial, because misspecified diffusion typically leads to poor drift estimation and vice-versa (Beven et al., 1993; Guan et al., 2025). This is problematic, since knowing an accurate estimate of drift or diffusion *a priori* is unrealistic in practice. Finally, we note that recent work estimates drift and diffusion, but does not consider the uniqueness of learned parameters, due to non-identifiability (Terpin et al., 2024; Brogat-Motte et al., 2024).

To advance the field of statistical inference from temporal marginals, we work to remove previous assumptions by determining conditions for identifiability and designing a method that leverages them. Wang et al. (2024) showed that identifiability of a linear additive noise SDE from ground truth trajectories with fixed $X_0 \in \mathbb{R}^d$ is equivalent to a non-degenerate rank condition based jointly on X_0, the linear drift A and the observational diffusion $GG^\top$. In contrast, our work studies identifiability of linear additive noise SDEs in the more general "marginals-only" setting. We also demonstrate that identifiability from continuous observation can be guaranteed across all linear additive noise SDEs solely by checking the initial distribution. To the best of our knowledge, this has not been shown even in the case of trajectory-based observations.

3. Identification of Linear SDEs from Temporal Marginals

In this section, we prove three identifiability results of increasing generality, which provide conditions for uniquely recovering the drift and diffusion of linear additive noise SDEs. Our theory demonstrates that SDE non-identifiability is only possible when the observed marginals share a generalized rotational symmetry. Furthermore, identifiability holds under generic conditions, since the set of SDE parameters that can preserve these symmetries over the course of the continuous evolution has Lebesgue measure 0.

3.1 SDE Identifiability Under Isotropic Diffusion

We first consider the case of SDEs with linear drift and isotropic diffusion:

$$dX_t = AX_t\mathrm{d}t + \sigma\mathrm{d}W_t, \tag{6}$$

and we consider the question of when we can jointly identify the drift A and diffusivity σ^2 from continuously observed marginals $(p_t : t \geq 0)$. It turns out that non-identifiability is intrinsically related to the marginals sharing a generalized rotational symmetry. We first define what we mean by a generalized rotation.

Definition 1 (Generalized rotation) *Let $d \geq 2$. We define a Σ-generalized rotation in $\mathbb{R}^d$ as a matrix exponential $e^{B\theta}$, such that θ is the rotation angle and $B \in \mathbb{R}^{d\times d}$ is skew-symmetric with respect to $\Sigma \succeq 0$, i.e. $B\Sigma + \Sigma B^\top = 0$.*

Note that classical rotations correspond to the case where we restrict $\Sigma = I_d$, since this enforces that the matrix B is skew-symmetric. By allowing $\Sigma \neq I_d$, we enable rotations to be defined over anisotropic geometries. If $\Sigma \succ 0$ is full rank, then Σ-generalized rotations are equivalent to classical rotations in a transformed space, after lengths are modified by scaling (see Lemma 14 in appendix). Otherwise, if $\Sigma \succeq 0$ is rank-deficient, then the condition $B\Sigma + \Sigma B^\top = 0$ forces B to leave the range of Σ invariant, and to be skew with respect to Σ there, while its action on the kernel of Σ is unconstrained. We now extend this notion to random variables, by defining a distribution X to be *auto-rotationally invariant* (ARI) if its law is invariant to the action $e^{B\theta}$ for some nonzero matrix $B \in \mathbb{R}^{d\times d}$.

Definition 2 (Auto-rotational invariance) *Let $d \geq 2$. We define a d-dimensional random variable X to be auto-rotationally invariant (ARI) if there exists some square matrix $B \neq 0$, such that for all $\theta \in \mathbb{R}$, $e^{B\theta}X \sim X$. For the case $d = 1$, we by convention consider mean-zero Gaussians $\mathcal{N}(0, \sigma^2)$ as the only ARI distributions.*

We use the term *auto-rotational* because it turns out that invariance of the law under the transformation $X \to e^{B\theta}X$ implies that $e^{B\theta}$ must be a generalized rotation (Definition 1), where the geometry Σ is the distribution's second moment $\Sigma = \mathbb{E}[XX^\top]$ (see Lemma 15). Auto-rotational invariance therefore requires that the law of the random variable is rotationally symmetric with respect to the geometry imposed by its own covariance structure. If $\Sigma = \mathbb{E}[XX^\top] \succ 0$ is full rank, then this is a highly restrictive property, since the density must be constant along orbits of $e^{B\theta}$, which each lie within an ellipsoid determined by Σ (see Lemma 16). For example, all mean zero Gaussians are auto-rotationally invariant (see Lemma 17). Otherwise, if $\Sigma = \mathbb{E}[XX^\top] \succeq 0$ is rank-degenerate, then the distribution also satisfies auto-rotational invariance (see Lemma 18).

We now present our first identifiability result, which states that the SDE parameters (A, σ^2) can be uniquely identified from continuously observed marginals $X_t \sim p_t$, unless each marginal is auto-rotationally invariant with respect to the same generalized rotation $e^{B\theta}$.

Theorem 3 (Identifiability under isotropic diffusion) *Suppose that the observed marginals $(p_t : t \geq 0)$ can be produced by multiple distinct parameter sets $(A, \sigma^2) \neq (\tilde{A}, \tilde{\sigma}^2)$. Then, each marginal $X_t \sim p_t$ is auto-rotationally invariant with respect to the same generalized rotational action $e^{B\theta}$, i.e. for all $X_t \sim p_t$, we have that $e^{B\theta} X_t \sim X_t$ for all $\theta \in \mathbb{R}$.*

Proof First, suppose that distinct parameter sets $(A, \sigma^2) \neq (\tilde{A}, \tilde{\sigma}^2)$ produce the same marginals $(p_t : t \geq 0)_{A,\sigma^2} = (p_t : t \geq 0)_{\tilde{A},\tilde{\sigma}^2}$. Then, their Fokker-Planck evolution equations are equal:

$$\begin{aligned}\partial_t p_t(x) &= -\nabla \cdot (Axp_t(x)) + \frac{\sigma^2}{2}\Delta p_t(x) \\ &= -\nabla \cdot (\tilde{A}xp_t(x)) + \frac{\tilde{\sigma}^2}{2}\Delta p_t(x).\end{aligned}$$

Without loss of generality, we may define the residual diffusivity $\bar{\sigma}^2 = \sigma^2 - \tilde{\sigma}^2 \geq 0$ and the residual drift $\bar{A} = A - \tilde{A}$. By subtracting the equations, we obtain a stationary Fokker-Planck equation with the residual parameters

$$0 = -\nabla \cdot (\bar{A}xp_t(x)) + \frac{\bar{\sigma}^2}{2}\Delta p_t(x). \tag{7}$$

First, we note that if the residual diffusion $\bar{\sigma}^2 = \sigma^2 - \tilde{\sigma}^2$ is nonzero, it follows from elliptic PDE theory that the residual Fokker-Planck equation (7) has at most one stationary distribution (Bogachev et al., 2022, Theorem 4.1.6, Example 4.1.8). Since each marginal p_t solves (7), each marginal is in fact the stationary distribution p_{eq} of a residual process with drift $\bar{A}$ and diffusivity $\bar{\sigma}^2$. Since the residual is a linear SDE, p_{eq} is a Gaussian distribution $\mathcal{N}(0, \Sigma)$ with Σ satisfying the Lyapunov equation $\Sigma\bar{A}^\top + \bar{A}\Sigma = -\bar{\sigma}^2 I_d$ (Gajic and Qureshi, 2008, Section 6.2.1). Thus, if the diffusivities are unequal, then non-identifiability implies that all marginals coincide at the same mean-zero Gaussian distribution. By Lemma 17, all such distributions satisfy auto-rotational invariance.

From this analysis, we may also deduce that if the marginals $(p_t)_{t\geq 0}$ are not all the same distribution, then non-identifiability is only possible if both SDEs share the same diffusivity σ^2, and are thus only distinct in their drifts. In this case, the residual Fokker-Planck equation reduces to the continuity equation:

$$0 = -\nabla \cdot (\bar{A}xp_t(x)). \tag{8}$$

By Liouville's theorem (see Neklyudov et al. (2021, Theorem 1)), each observed marginal $X_t \sim p_t$ is stationary under the linear ODE $\dot{x} = \bar{A}x$. Thus, for each marginal X_t, we have

$$e^{\bar{A}\theta} X_t \sim X_t \quad \forall \theta \in \mathbb{R}. \tag{9}$$

We conclude that each marginal $X_t \sim p_t$ is auto-rotationally invariant with respect to the same matrix $B = \bar{A}$. ■

We make a couple of remarks about this result. First, we recall that observing marginals at equilibrium is a well-documented source of non-identifiability, since equilibrium only depends on the drift to diffusion ratio (see Lavenant et al. (2021) and Guan et al. (2025, Proposition 4.1)). Our result shows that under isotropic diffusion, observing marginals that vary in time is enough to determine the true diffusivity σ^2. This follows from the same elliptic PDE argument underlying a recent identifiability result by Guan et al. (2025), which showed that observation at equilibrium is the only source of non-identifiability of the SDE parameters for gradient-flow SDEs, which are defined by conservative drift $-\nabla\Psi$ and isotropic diffusivity σ^2.

However, in contrast to the gradient-flow setting, linear SDEs may have rotational drift, which introduces undetectable rotation as a second source of non-identifiability. Indeed, a classic non-identifiability example from the literature is given by the systems (Hashimoto et al., 2016; Shen et al., 2024):

$$\mathrm{d}X_t = \mathrm{d}W_t, \qquad X_0 \sim \delta_0 \tag{10}$$

$$\mathrm{d}Y_t = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} Y_t \,\mathrm{d}t + \mathrm{d}W_t, \qquad Y_0 \sim \delta_0, \tag{11}$$

where the residual drift is the skew-symmetric matrix $\bar{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, which produces a rotational action that is indistinguishable at the level of the marginals $p_t \sim \mathcal{N}(0, tI_d)$. We present examples with more general rotational geometries in Section B. Our result shows that this failure mode, characterized by shared invariance under a generalized rotational action, is the only other source of non-identifiability besides observation at equilibrium, for linear SDEs with isotropic diffusion.

3.2 SDE Identifiability Under Anisotropic Diffusion

We now consider the more general case of linear additive noise SDEs with possibly anisotropic diffusion:

$$dX_t = AX_t\mathrm{d}t + G\mathrm{d}W_t, \tag{12}$$

and we derive identifiability conditions for determining the SDE parameters $(A, H = GG^\top)$ from observed marginals $(p_t : t \geq 0)$. We note that anisotropic diffusion significantly complicates identifiability analysis, since the diffusion level can vary in each dimension. Indeed, anisotropic diffusion can produce non-identifiability such that the distributions are stationary in a subset of components, while otherwise exhibiting transience. Since equilibrium distributions for linear additive noise are mean-zero Gaussians (Pavliotis, 2014, Sections 4.2 and 4.4), we can illustrate this by considering a 'Gaussian pancake' initial distribution $X_0 \sim p_0$, where $X_0^{(1)} \sim \mathcal{N}(0,1)$ and the remaining components $X_0^{(2:d)}$ are initialized independently. To produce non-identifiability, we consider two different univariate dynamics that preserve $X_t^{(1)} \sim \mathcal{N}(0,1)$ and we decouple these first component dynamics from the rest

of the system:

$$\mathrm{d}X_t = \begin{bmatrix} -\frac{1}{2} & \mathbf{0} \\ \mathbf{0} & A_{\setminus 1} \end{bmatrix} X_t \,\mathrm{d}t + \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & G_{\setminus 1} \end{bmatrix} \mathrm{d}W_t, \quad X_0 \sim p_0 \tag{13}$$

$$\mathrm{d}Y_t = \begin{bmatrix} -2 & \mathbf{0} \\ \mathbf{0} & A_{\setminus 1} \end{bmatrix} Y_t \,\mathrm{d}t + \begin{bmatrix} 2 & \mathbf{0} \\ \mathbf{0} & G_{\setminus 1} \end{bmatrix} \mathrm{d}W_t, \quad Y_0 \sim p_0. \tag{14}$$

These SDEs are non-identifiable from marginals since they are both stationary in the first component, while being trivially identical in other components since they share the same independent complementary dynamics $(A_{\setminus 1}, G_{\setminus 1})$.

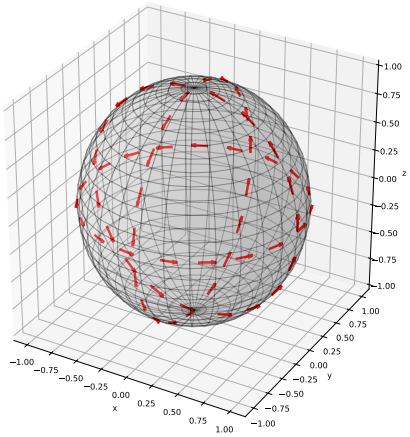
While we saw that Gaussianity is directly related to non-identifiability in the previous setting with isotropic diffusion, the above construction shows that the addition of Gaussian slices can lead to non-identifiability if we permit anisotropic diffusion. Thus, non-identifiability is feasible even if a marginal $X_t \sim p_t$ fails to be auto-rotationally invariant, as defined in Definition 2. However, by recalling the classical fact that Gaussians are the only random variables that have finitely many nonzero cumulants (Marcinkiewicz, 1939), and moreover, that the only nonzero cumulant of a mean-zero Gaussian is its covariance, we may consider a natural relaxation of auto-rotational invariance. Instead of defining the invariance relation over the full law, we can define it on all cumulant orders, except $n = 2$.

Definition 4 (Weak auto-rotational invariance) *Let X be a d-dimensional r.v, and denote by $T_n^X(v, \ldots, v)$ the order-n tensor defining the nth cumulant of X, projected in the direction of $v \in \mathbb{R}^d$, i.e. the nth cumulant of the univariate r.v. $v^\top X$. Then, X is weakly auto-rotationally invariant if there exists some nonzero matrix $B \in \mathbb{R}^{d\times d}$, such that the mean and higher order cumulants $n \geq 3$ obey the invariance:*

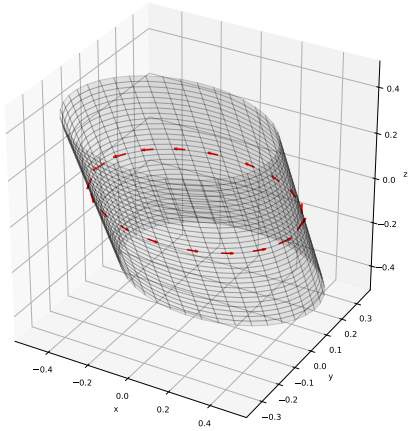
$$T_n^X(e^{B^\top \theta} v, \ldots, e^{B^\top \theta} v) = T_n^X(v, \ldots, v) \quad \forall n \in \mathbb{N} \setminus \{2\}, \forall v \in \mathbb{R}^d, \forall \theta \in \mathbb{R}. \tag{15}$$

For convenience, we sometimes drop the "weak" qualifier and simply refer to weak ARI as ARI, but we make the distinction explicit in formal statements and proofs. Indeed, as shown in Lemma 19, if the support of the random variable is bounded, then weak ARI (Definition 4) is in fact equivalent to full ARI (Definition 2). Intuitively, weak ARI omits the constraint at order $n = 2$, which enables the addition of Gaussian slices. While we saw that degenerate distributions, which are supported in a proper subspace of $\mathbb{R}^d$ are always ARI (and hence weakly ARI), weak ARI also encompasses random variables that are Gaussian in a proper subspace of $\mathbb{R}^d$. In Figure 2, we visualize three ARI distributions as well as a Gaussian pancake distribution, which is only weakly ARI.

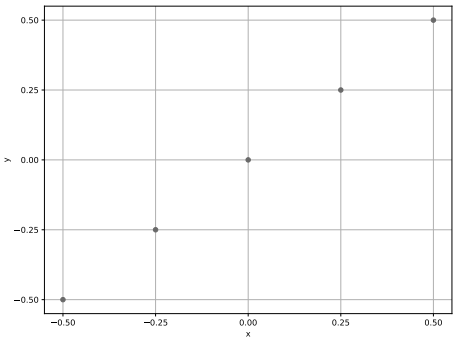
We now present our second identifiability theorem, which proves that non-identifiability from the marginals of a linear additive noise SDE is only possible if the initial marginal $X_0 \sim p_0$ is weakly auto-rotationally invariant, and furthermore, that there exist non-identifiable systems for any weakly ARI initial distribution X_0. Intuitively, non-identifiability is only feasible if there is a symmetry to exploit, and if the initial distribution has such a symmetry, then it is susceptible to non-identifiability.



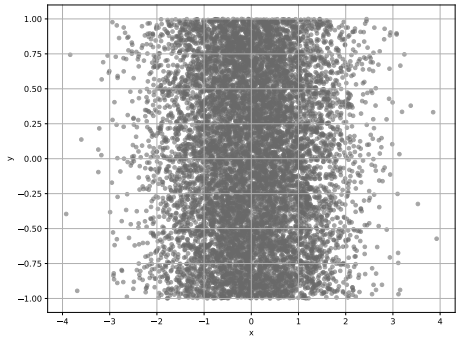
(a) X_0 is spherical and invariant under all classical rotations.



(b) X_0 is cylindrical (up to scaling), and is invariant under rotation in the xy plane after transformation.



(c) X_0 is degenerate and its support lies in the nullspace of $B = \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}$



(d) X_0 is a Gaussian pancake, which is mean 0 univariate Gaussian in x.

Figure 2: We visualize four distributions that satisfy auto-rotational invariance. The first three distributions obey strong auto-rotational invariance (Definition 2), while the Gaussian pancake is weakly auto-rotationally invariant (Definition 4).

Theorem 5 (Identifiability for linear additive noise SDEs) *Let p_0 be a probability distribution on $\mathbb{R}^d$ with a well-defined moment generating function, and let*

$$\Theta = \{(A, H) : A \in \mathbb{R}^{d\times d}, G \in \mathbb{R}^{d\times m}, H = GG^\top\}$$

denote the set of drift-diffusion pairs under a linear additive noise model (3)*. It follows that every parameter configuration $(A, H) \in \Theta$ induces a unique set of marginals $(p_t : t \geq 0)_{A,H}$ from initialization $X_0 \sim p_0$, if and only if X_0 is not weakly auto-rotationally invariant (Definition 4).*

Proof We first prove that non-identifiability can only occur if X_0 is weakly ARI. Suppose that two linear additive noise SDEs have distinct parameters $(A, H = GG^\top) \neq (\tilde{A}, \tilde{H} =$

$\tilde{G}\tilde{G}^\top$), yet equal marginals for all times $t \geq 0$. We note that these have analytic solutions (see Section A):

$$X_t = e^{At}X_0 + \int_0^t e^{A(t-s)}GdW_s$$
$$Y_t = e^{\tilde{A}t}X_0 + \int_0^t e^{\tilde{A}(t-s)}\tilde{G}dW_s,$$

where the initial distribution X_0 is shared.

Then, since the m.g.f is well-defined, equality in distribution $X_t \sim Y_t$ is equivalent to equality in all cumulants between X_t and Y_t:

$$T_n^{X_t}(v,\dots,v) = T_n^{Y_t}(v,\dots,v) \qquad \forall v \in \mathbb{R}^d, \forall t \geq 0, \tag{16}$$

where $T_n^X(v,\dots,v)$ denotes the nth cumulant of X, projected onto $v \in \mathbb{R}^d$. By matching cumulants of all orders, non-identifiability requires that the following infinite system of equations is satisfied for all $v \in \mathbb{R}^d$ and $t \geq 0$:

$$\begin{cases} T_n^{X_0}(e^{A^\top t}v,\dots,e^{A^\top t}v) & = T_n^{X_0}(e^{\tilde{A}^\top t}v,\dots,e^{\tilde{A}^\top t}v) \qquad n \neq 2, \\ T_2^{X_0}(e^{A^\top t}v, e^{A^\top t}v) + v^\top \Sigma_t v & = T_2^{X_0}(e^{\tilde{A}^\top t}v, e^{\tilde{A}^\top t}v) + v^\top \tilde{\Sigma}_t v \qquad n = 2, \end{cases} \tag{17}$$

Note that the Ito integral terms contribute only to the second order cumulant, since they are mean zero Gaussians (Marcinkiewicz, 1939). We are also able to move the matrix exponential into the arguments, due to the homogeneity of cumulant tensors. We first focus on all cumulants of order $n \neq 2$ and note that they imply the identity:

$$T_n^{X_0}(e^{\bar{A}^\top t}v,\dots,e^{\bar{A}^\top t}v) = T_n^{X_0}(v,\dots,v) \qquad n \neq 2 \tag{18}$$

Indeed, if we take the time derivative at $t = 0$, we obtain

$$\begin{aligned} \left.\frac{d}{dt}\right|_{t=0} T_n^{X_0}(e^{A^\top t}v,\dots,e^{A^\top t}v) &= \left.\frac{d}{dt}\right|_{t=0} T_n^{X_0}(e^{\tilde{A}^\top t}v,\dots,e^{\tilde{A}^\top t}v) \\ \implies nT_n^{X_0}(A^\top v, v,\dots,v) &= nT_n^{X_0}(\tilde{A}^\top v, v,\dots,v) \\ \implies T_n^{X_0}(\bar{A}^\top v, v,\dots,v) &= 0 \end{aligned}$$

where $\bar{A} = A - \tilde{A}$. Since the above holds for all $v \in \mathbb{R}^d$, we can apply it to any vector in the set $\{u(t) = e^{\bar{A}^\top t}v : t \geq 0, v \in \mathbb{R}^d\}$, which yields

$$\begin{aligned} & T_n^{X_0}(\bar{A}^\top u(t), u(t),\dots,u(t)) = 0 \\ \implies \frac{d}{dt}T_n^{X_0}(u(t),\dots,u(t)) &= nT_n^{X_0}(\bar{A}^\top u(t), u(t),\dots,u(t)) = 0. \end{aligned}$$

Thus, along the trajectory $u(t) = e^{\bar{A}^\top t}v$, for all $n \neq 2$, the nth cumulant of X_0, projected onto $u(t)$, is constant in time, so the equations for matching the higher order cumulants reduce to

$$T_n^{X_0}(e^{\bar{A}^\top t}v,\dots,e^{\bar{A}^\top t}v) = T_n^{X_0}(v,\dots,v) \qquad \forall n \neq 2,$$

which is precisely the condition for weak auto-rotational invariance (15) with $B = \bar{A}$.

We now show that if X_0 satisfies the weak ARI condition (15) for some matrix $B \neq 0$, then there exist systems with distinct parameter sets $(A, H = GG^\top) \neq (\tilde{A}, \tilde{H} = \tilde{G}\tilde{G}^\top)$, which produce the same marginals for each time, i.e. $(p_t : t \geq 0)_{A,H} = (p_t : t \geq 0)_{\tilde{A},\tilde{H}}$. We denote the initial second moment $\Sigma = \mathbb{E}[X_0 X_0^\top]$, and consider the rank-degenerate and full rank cases separately.

First, suppose Σ is rank-degenerate, so that some $u \neq 0$ has $\Sigma u = 0$ and hence $\mathbb{E}[(u^\top X_0)^2] = u^\top \Sigma u = 0$, giving $u^\top X_0 = 0$ a.s. For our non-identifiable counterexample, we consider base parameters $A = -\alpha I$ and $H = 2\alpha\Sigma$. The idea is that since the initial distribution and the noise are supported in the hyperplane $u^\perp$ ($\text{ran}G = \text{ran}H = \text{ran}\Sigma \subseteq u^\perp$), the marginals will always stay in this subspace. This allows us to alter the drift in the direction of u, without changing the evolution.

In particular, we can pick any $w \neq 0$ and set the drift offset $\bar{A} = A - \tilde{A} = wu^\top \neq 0$, so that $\ker \bar{A} = u^\perp$, and take the same diffusion $\tilde{G} = G \implies \tilde{H} = H$. Then $\bar{A}X_0 = w(u^\top X_0) = 0$ a.s. and $\bar{A}G = 0$. Then, since $-\alpha I$ commutes with $\bar{A}$, expanding the exponential $e^{\tilde{A}t}X_0$ gives $e^{\tilde{A}t}X_0 = e^{-\alpha t}X_0$ a.s. Similarly, $e^{\tilde{A}(t-s)}G = e^{-\alpha(t-s)}G$. Driven by the same Brownian motion, the two systems therefore produce identical paths, despite having a drift offset of $\bar{A}$. We provide one such construction in Example 3.

To design a non-identifiable system in the non-degenerate case, where $\Sigma \succ 0$ is full rank, we consider an SDE with base parameters

$$A = -\alpha I$$
$$H = 2\alpha\Sigma$$

Then, to define the second system, we set the drift difference to be the matrix for which X_0 is weakly ARI, i.e. $\bar{A} = A - \tilde{A} = B$, such that the second system has parameters

$$\tilde{A} = A - \bar{A} = -\alpha I - \bar{A}$$
$$\tilde{H} = H - \bar{H} = 2\alpha\Sigma + \bar{A}\Sigma + \Sigma\bar{A}^\top$$

Note that since $\Sigma \succ 0$, then we can pick α arbitrarily large to ensure that $\tilde{H}$ has positive spectrum, and is a valid p.s.d diffusion matrix. Since $A = -\alpha I$ is isotropic, it follows that we have the analytic solutions

$$X_t = e^{-\alpha t}X_0 + \mathcal{N}(0, \Sigma_t) \qquad \text{s.t. } \Sigma_t = \int_0^t e^{A(t-s)} H e^{A^\top (t-s)} ds$$
$$Y_t = e^{-\alpha t}e^{-\bar{A}t}X_0 + \mathcal{N}(0, \tilde{\Sigma}_t) \qquad \text{s.t. } \tilde{\Sigma}_t = \int_0^t e^{\tilde{A}(t-s)} \tilde{H} e^{\tilde{A}^\top (t-s)} ds.$$

Then, note that non-identifiability holds if and only if $X_t \sim Y_t$ for all $t \geq 0$. Equivalently, the cumulants of all orders n match for all times and projection vectors $v \in \mathbb{R}^d$:

$$T_n^{(X_t)}(v, \ldots, v) = T_n^{(Y_t)}(v, \ldots, v) \qquad \forall n \in \mathbb{N}, \forall v \in \mathbb{R}^d, \forall t \geq 0 \tag{19}$$

To show that the above holds, we consider the cases $n \neq 2$ and $n = 2$ separately. Matching all cumulant orders $n \neq 2$ reduces to showing that $e^{-\alpha t}X_0$ and $e^{-\alpha t}e^{-\bar{A}t}X_0$ have the same

higher order cumulants. This follows directly from X_0 being weakly ARI with respect to $\bar{A}$:

$$\begin{aligned} T_n^{e^{-\alpha t}X_0}(v,\dots,v) &= e^{-n\alpha t} T_n^{X_0}(v,\dots,v) \\ &\overset{(15)}{=} e^{-n\alpha t} T_n^{X_0}(e^{-\bar{A}^\top t}v,\dots,e^{-\bar{A}^\top t}v) \\ &= T_n^{e^{-\alpha t}e^{-\bar{A}t}X_0}(v,\dots,v). \end{aligned}$$

Then, to show that the cumulants match for order $n = 2$, we note that this corresponds to equality in covariance. Since the case $n \neq 2$ above includes $n = 1$, the two systems already share the same mean at all times, so it suffices to show that their second moments agree. We in fact constructed the systems so that the second moment is constant in time, i.e. $\mathbb{E}[X_t X_t^\top] = \Sigma = \mathbb{E}[Y_t Y_t^\top]$ for all $t \geq 0$. Indeed, the second moment of a linear additive noise SDE solves the autonomous Lyapunov ODE $\dot{M}_t = AM_t + M_t A^\top + H$, so it suffices to check that $M_t = \Sigma$ is an equilibrium solution. With $A = -\alpha I$, $H = 2\alpha\Sigma$, we obtain

$$\begin{aligned} A\Sigma + \Sigma A^\top + H &= -2\alpha\Sigma + 2\alpha\Sigma = 0 \\ \tilde{A}\Sigma + \Sigma\tilde{A}^\top + \tilde{H} &= (A\Sigma + \Sigma A^\top + H) - \left(\bar{A}\Sigma + \Sigma\bar{A}^\top + \bar{H}\right) \\ &= -\left(\bar{A}\Sigma + \Sigma\bar{A}^\top + \bar{H}\right) = 0. \end{aligned}$$

Thus, all cumulants match between both SDEs at all times, and we conclude that they are non-identifiable. ■

Theorem 5 shows that weak auto-rotational invariance of the initial distribution is a necessary and sufficient condition for non-identifiability to be feasible. We note that by re-indexing the index of the initial time, it follows that identifiability of the SDE parameters is guaranteed if any of the marginals in the evolution is not weakly ARI. Furthermore, since the SDE parameters are constant in time, non-identifiability not only implies that each marginal is weakly ARI, but also that each marginal is weakly auto-rotationally invariant with respect to the same matrix B in the condition (15). In other words, non-identifiability requires that a symmetry invariance is shared across all marginals. We now determine the prevalence of these shared symmetries, in order to quantify the risk of non-identifiability. Below, we prove that the SDE parameters are generically identifiable, even if X_0 is weakly ARI. In particular, we show that the curve of covariances $(\mathrm{Cov}(X_t) : t \geq 0)$ uniquely identifies the true SDE parameters (A, H) for almost every parameter configuration.

Theorem 6 (Generic identifiability for linear additive noise SDEs) *Suppose $C_0 = \mathrm{Cov}(X_0)$ is the covariance of an initial distribution $X_0 \sim p_0$. Then, for Lebesgue-almost every $(A, G) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times m}$ under the linear additive noise model, the covariance curve $(C_t : t \geq 0)_{A,G}$ uniquely determines the SDE parameters $(A, H = GG^\top)$. Thus, the SDE parameters (A, H) are generically identifiable under continuous observation.*

Proof We wish to prove that the map $(A, G) \to (C_t : t \geq 0)_{A,G}$ uniquely determines $(A, H = GG^\top)$ on $(\mathbb{R}^{d\times d} \times \mathbb{R}^{d\times m}) \setminus \mathcal{N}$, where $\mathcal{N}$ has Lebesgue measure 0. First, we note that if two different parameter configurations trace out the same curve of covariances, i.e.

$(C_t : t \geq 0)_{A,G} = (C_t : t \geq 0)_{\tilde{A},\tilde{G}}$, then by Lemma 21, their residual parameters $\bar{A} = A - \tilde{A}$ and $\bar{H} = GG^\top - \tilde{G}\tilde{G}^\top$ satisfy the Lyapunov equation

$$\bar{A}C_t + C_t\bar{A}^\top = -\bar{H}. \tag{20}$$

Thus, if we consider a collection of D nonzero times $\{t_i\}_{i=1}^D$, and define $\Delta C_i = C_{t_i} - C_0 = \mathrm{Cov}(X_{t_i}) - \mathrm{Cov}(X_0)$, then we can subtract the above equation at times $t = t_i$ and $t = 0$ to obtain

$$\bar{A}\Delta C_i + \Delta C_i\bar{A}^\top = 0 \qquad \forall i = 1, \ldots, D. \tag{21}$$

The main fact that we would like to show is that up to a set $\mathcal{N}$ of measure 0, the condition $(C_t : t \geq 0)_{A,G} = (C_t : t \geq 0)_{\tilde{A},\tilde{G}}$ forces the residual drift $\bar{A}$ to be 0. From this, the drifts would be equal, and it would follow that the corresponding diffusion matrices $H = GG^\top$ and $\tilde{H} = \tilde{G}\tilde{G}^\top$ must also be equal, since otherwise, the curve of covariances would deviate (see equations (50) and (51) in the proof of Lemma 21).

To prove the above fact, it suffices to show that the covariance differences $\{\Delta C_i\}_{i=1}^D$ span Sym_d, since it would then follow that $\bar{A} = 0$ is the only solution to (21). Indeed, assume that the covariance differences span Sym_d. Then, since $\mathrm{I}_d \in \mathrm{Sym}_d$, there exists a linear combination of (21), which produces $\bar{A} + \bar{A}^\top = 0$, i.e., $\bar{A}$ is skew-symmetric. Thus, we can rewrite (21) as $\bar{A}\Delta C_i = \Delta C_i\bar{A}$. Then, since by hypothesis, every diagonal matrix is in the span of the covariance differences, it follows that $\bar{A}$ commutes with every diagonal matrix. Taking $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ with distinct entries, the identity $\bar{A}\Lambda = \Lambda\bar{A}$ reads $\bar{A}_{ij}\lambda_j = \lambda_i\bar{A}_{ij}$ entrywise, forcing $\bar{A}_{ij} = 0$ whenever $i \neq j$. Hence $\bar{A}$ is diagonal, and therefore symmetric. However, $\bar{A} = 0$ is the only matrix that is both symmetric and skew-symmetric.

We now show that the covariance differences span Sym_d. For this, note that Sym_d is a $\frac{d(d+1)}{2}$ dimensional real subspace, since symmetric matrices are determined by the upper triangular block (with main diagonal). We therefore consider $D = \frac{d(d+1)}{2}$ covariance differences and apply the half-vectorization operator vech, which stacks the entries on and below the main diagonal of a symmetric matrix. We then form the $D \times D$ matrix

$$M(A, G) = \begin{bmatrix} \mathrm{vech}(\Delta C_1)^\top \\ \vdots \\ \mathrm{vech}(\Delta C_D)^\top \end{bmatrix}.$$

Indeed, without loss of generality, we may regard the covariance differences with respect to C_0 as functions of (A, G). To conclude the proof, we show that $\det(M(A, G)) \neq 0$ for Lebesgue-almost every $(A, G) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times m}$. To do so, we use the fact that a real analytic function is either identically zero or nonzero Lebesgue-almost everywhere (Mityagin, 2015, Proposition 0), and apply this to $\det(M(A, G))$. Indeed,

$$C_t = e^{At}C_0e^{A^\top t} + \int_0^t e^{As}GG^\top e^{A^\top s}\, ds$$

is analytic in (A, G), so it follows that the determinant of $M(A, G)$ is also analytic. We therefore just need to show that for a given initial covariance C_0, there exists a single

parameter configuration (A, G) such that $\det(M(A, G)) \neq 0$. We provide an example in Proposition 22 of the Appendix.

■

To provide some intuition, we note that the converse direction of Theorem 5 was shown by adversarially choosing SDE parameters that leave the covariance stationary, thus enabling non-identifiability by fixing the geometry across all marginals. However, even if the initial distribution X_0 is weakly ARI, Theorem 6 shows that almost every parameter configuration will produce an evolution that breaks the symmetries required for non-identifiability.

We conclude with a few remarks about the practical implications of our identifiability results and connections to related theory. First, we note the limitation that our identifiability results assume continuous observation of the marginals $X_t \sim p_t$. We note that the assumption of continuous observation cannot generally be relaxed, due to the aliasing problem. For example, let $p_0 = \text{Unif}[-1, 1]^2$ be a uniform square, which is clearly not auto-rotationally invariant. If we consider a rotational action from the skew-symmetric drift $A = \begin{bmatrix} 0 & \gamma \\ -\gamma & 0 \end{bmatrix}$ and $H = 0$, and only sample the marginals at discrete times $\{0, \Delta t, \cdots, k\Delta t, \cdots\}$, then it would be impossible to determine whether the square has rotated by θ in between two times, or by $2\pi k + \theta$ for any $k \in \mathbb{N}$. This is precisely the Nyquist aliasing phenomenon in signal processing: discrete sampling at rate $\frac{1}{\Delta t}$ cannot resolve frequencies above $\frac{\pi}{\Delta t}$. However, we note that aliasing counterexamples require a strict synchronicity, and moreover that generic SDE parameter configurations will also alter the covariance geometry, as we showed in Theorem 6. We therefore expect that identifiability results can be proven in the finite-sample setting under certain assumptions, but note that new techniques must be used to prove them, since we cannot directly leverage the differential structure of the evolution equations.

Second, by proving that non-identifiability requires auto-rotational invariance, which defines a shared invariance structure on the cumulants (15), our results show that higher order cumulants provide crucial information for identifiability. This bears interesting parallels to identifiability theory in causal discovery for linear non-Gaussian acyclic models (LiNGAM) on static data, where the presence of higher order cumulants from non-Gaussian noise enables the causal graph to be identified in the absence of latent confounders (Shimizu et al., 2006). In each of these settings, we seek identifiability from data composed of linear effects with additive independent noise. In the static case, we are unable to resolve cycles in the graph, and imposing non-Gaussianity in the independent noise provides sufficient cumulant information to identify the DAG. Meanwhile, for linear SDEs, the temporal component allows us to identify cycles in the graph. Indeed, as we will see in the next section, the drift matrix (A) trivially determines all directed edges in the endogenous set. However, the noise is necessarily Gaussian, since the diffusion is driven by Brownian motion. Thus, identifiability in both settings is related to higher order cumulants. In the static setting, the source of the higher order cumulants comes from non-Gaussian noise, whereas in the SDE setting, the Brownian noise is Gaussian, and thus, the higher order cumulants come from the initial distribution.

4. Causal Graph Identification from SDE Parameters

In this section, we establish theory for recovering the causal graph of an additive noise SDE from its identified drift and (observational) diffusion parameters. Thus, given our identifiability results from the previous section, these insights demonstrate that under mild conditions, we can theoretically identify the SDE and the causal graph of a system that is governed by a linear additive noise SDE, directly from its temporal marginals $(p_t)_{t\geq 0}$.

First, we review the definition of a dynamic structural causal model (DSCM), which we adapt from Boeken and Mooij (2024, Definition 1).

Definition 7 (DSCM) *Let V be a set of d endogenous variables $\{X^{(1)}, \dots, X^{(d)}\}$ and let W be a set of exogenous variables, $\{X^{(d+1)}, \dots, X^{(d+m)}\}$, e.g. noise sources that influence the endogenous variables. Then, a DSCM is defined for each of the d endogenous variables, such that*

$$X_t^{(j)} = X_0^{(j)} + \int_0^t a_j(s, X_s^{\alpha(j)})\, \mathrm{d}h_j(s, X_s^{\beta(j)}), \tag{22}$$

where $\alpha(j) \subset V \cup W$ represents the variables used as arguments for the integrand $a_j(t, X_t)$, and $\beta(j) \subset V \cup W$ represents the variables used as arguments for the integrator $h_j(t, X_t)$. For a d-dimensional additive noise SDE (2) driven by an m-dimensional Brownian motion, the exogenous set is simply $W = \{X^{(d+1)}, \dots, X^{(d+m)}\} = \{W^{(1)}, \dots, W^{(m)}\}$, and the DSCM is defined by

$$X_t^{(j)} = X_0^{(j)} + \int_0^t b_j(X_s^{\alpha(j)})\, \mathrm{d}s + \sum_{k=1}^{m} G_{j,k} W_t^{(k)}. \tag{23}$$

In particular, $\alpha(j) \subset V$ depends only on the endogenous variables, and $\beta(j)$ corresponds to the components k of the Brownian motion (index $k+d$ in the expanded vector X), such that $G_{j,k} \neq 0$.

We can then define a causal graph $\mathcal{G}$ from the DSCM. We adapt (Boeken and Mooij, 2024, Definition 2), with the only modification being that we permit self-edges $j \to j$.

Definition 8 (Causal graph from DSCM) *Given a DSCM (22), we define a causal graph $\mathcal{G}$ with vertices $V \cup W$, and edges $E = \{i \to j : j \in V, i \in \alpha(j) \cup \beta(j)\}$.*

Intuitively, an edge $i \to j$ indicates that component $X_t^{(i)}$ directly influences the evolution of component $X_t^{(j)}$, and we only include edges that point towards variables in the endogenous set. Since the diffusion G is state-independent for additive noise SDEs, this means that endogenous variables can only influence one another through the drift function b. However, the diffusion still determines incoming edges to endogenous variables. In particular, latent confounders are modeled by off-diagonal entries of the diffusion G, since multiple variables would be correlated due to shared driving noise (see (Boeken and Mooij, 2024, Figure 1)).

This causal framework is well-defined given access to the drift b and diffusion G. However, due to the symmetry of Brownian motion, one can at best identify $H = GG^\top$ from observational data. Given the parameters b and $H = GG^\top$ from an additive noise SDE, we show that one can recover all endogenous edges from the drift b and also use non-zero entries in the observational diffusion $H = GG^\top$ to determine pairwise latent confounders.

Lemma 9 *Let X_t evolve according to a d-dimensional additive noise SDE: $\mathrm{d}X_t = b(X_t)\mathrm{d}t + G\mathrm{d}W_t$, and let $\mathcal{G} = (V \cup W, E)$ be its causal graph.*

a. There exists a directed edge $i \to j$ in $\mathcal{G}$ if and only if $b_j(X_t)$ depends on $X_t^{(i)}$.

b. There exists a latent confounder $W^{(k)}$ over (i, j) in $\mathcal{G}$, if $H_{i,j} = (GG^\top)_{i,j} \neq 0$.

Proof The first claim follows immediately from the definition of the dynamic structural causal model (23). Then, to prove the second claim, we note that if $H_{i,j} = G_{i,\cdot} \cdot G_{\cdot,j}^\top \neq 0$, then it follows that for some column $k \in [m]$ in the diffusion matrix G, we have that $G_{i,k}, G_{j,k} \neq 0$. From this, we conclude that $X^{(i)}$ and $X^{(j)}$ both share a dependence on the noise from the kth component of the Brownian motion, $W_t^{(k)}$.

■

Lemma 9 establishes a bridge between causal graph identification and SDE parameter identification, where we can only hope to recover the observational diffusion. Indeed, by combining Theorem 6 and Lemma 9, we see that we can almost always identify all directed edges within the endogenous set (due to recovery of the drift A) from the marginals of a linear additive noise SDE. Furthermore, the nonzero entries from the identified diffusion matrix, $H_{i,j}$, $i \neq j$, always identify pairwise latent confounding from shared noise. We emphasize that even when the drift is assumed to be linear, additive noise SDEs can encode arbitrary causal structures between endogenous variables (ex. via nonzero entries in the drift A). This makes our identification theory particularly relevant for GRN inference, since our setting can model cyclic behaviour, such as feedback loops and repressilator dynamics, which are fundamentally excluded in common gradient-flow models (Weinreb et al., 2018; Hashimoto et al., 2016; Lavenant et al., 2021; Terpin et al., 2024; Guan et al., 2025). We however note that not all noise relations can be learned from the observational diffusion, as we discuss in Examples 4 and 5 in Appendix D. We include additional details and discussion about the causal interpretation of SDEs in Appendix D.

5. Our Parameter Estimation Method

In this section, we introduce our method, Alternating Projection Parameter Estimation from X_0 (APPEX), which estimates a process' drift, diffusion, and causal graph from observed temporal marginals. Figure 3 illustrates the general strategy of the APPEX algorithm. We focus on linear additive noise SDEs, since Theorem 5 guarantees identifiability given that X_0 is not auto-rotationally invariant. However, our algorithm can be used to estimate any additive noise SDE in the form of Equation 2. In this section, we give an intuitive explanation at the population level, using the process' true temporal marginals $X_{t_i} \sim p_i$. In Section 6, we perform experiments using empirical marginals $\hat{p}_i$, with finite samples per marginal.

Let $p_0, ..., p_{N-1}$ denote a set of observed temporal marginals, and let $(A^{(0)}, H^{(0)})$ denote initial guesses for the drift and diffusion. The idea is to use an alternating optimization algorithm to obtain increasingly better estimates for the drift and diffusion matrices. Formally,

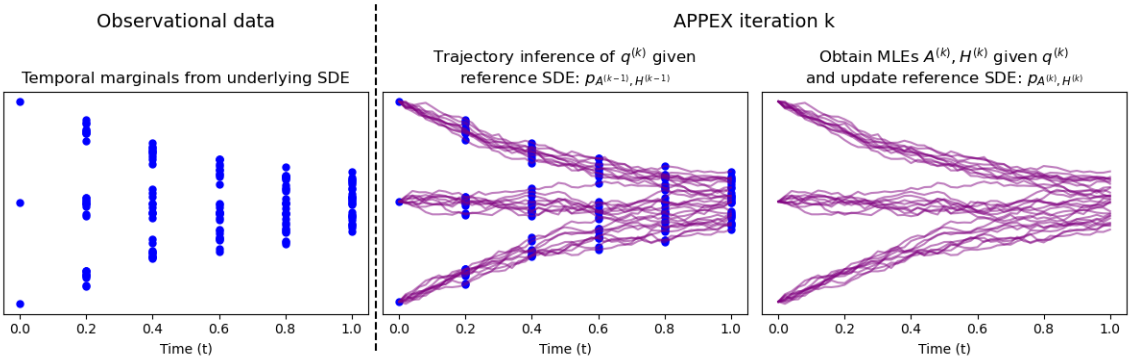


Figure 3: Visualization of our parameter estimation algorithm APPEX on a toy example. Given observed temporal marginals from an underlying SDE (left), APPEX alternates between trajectory inference (middle) and MLE parameter estimation (right) in order to find the SDE parameters that best represent the temporal marginal observations.

we consider the spaces

$$\mathcal{D} = \{q : q_0 = p_0, ..., q_{N-1} = p_{N-1}\} \text{ and} \tag{24}$$

$$\mathcal{A} = \left\{ p : \exists\, (A, H) \in \mathbb{R}^{d\times d} \times S^d_{\succeq 0} \text{ s.t } p(x, t+\Delta t \mid y, t) = \mathcal{N}\Big(e^{A\Delta t} y,\ \int_0^{\Delta t} e^{As}\, H\, e^{A^\top s}\, ds\Big) \right\}. \tag{25}$$

$\mathcal{D}$ is the set of laws on paths, which coincide with the N temporal marginals at their respective times. $\mathcal{A}$ is the set of laws given by linear additive noise SDEs. To find the underlying linear additive noise SDE in the intersection $\mathcal{A} \cap \mathcal{D}$, our method alternates between information projections, $\arg\min_{q\in\mathcal{D}} D_{KL}(q\|p)$ and moment projections, $\arg\min_{p\in\mathcal{A}} D_{KL}(q\|p)$. The information projection step corresponds to trajectory inference, whose goal is to find the law on paths $q \in \mathcal{D}$, which best aligns with the reference measure p. The moment projection then corresponds to maximum likelihood parameter estimation, since it amounts to determining the law on paths $p \in \mathcal{A}$ over linear additive noise SDEs, which best aligns with the law of the estimated trajectories q. Indeed, in the infinite-data limit, MLE is equivalent to minimizing relative entropy, i.e. KL divergence (Akaike, 1998; White, 1982). At the population level, an iteration entails the updates:

$$q^{(k)} = \arg\min_{q\in\mathcal{D}} D_{KL}(q\|p_{A^{(k-1)},H^{(k-1)}}) \tag{26}$$

$$A^{(k)} = \arg\min_{A\in\mathbb{R}^{d\times d}} D_{KL}(q^{(k)}\|p_{A,H^{(k-1)}}) \tag{27}$$

$$H^{(k)} = \arg\min_{H\in S^d_{\succeq 0}} D_{KL}(q^{(k)}\|p_{A^{(k)},H}). \tag{28}$$

At iteration k, we start by using the SDE parameters estimated from the previous iteration to define a reference SDE $p_{A^{(k-1)},H^{(k-1)}} \in \mathcal{A}$. Using this reference SDE and the set of observed temporal marginals, we perform trajectory inference (26), to estimate the path

distribution $q^{(k)} \in \mathcal{D}$ that minimizes relative entropy to the reference SDE, while satisfying the marginal constraints (24) (see the middle plot of Figure 3). We then perform a two-step maximum likelihood estimation, in order to estimate the drift and diffusion from the estimated trajectories. This in turn updates the reference SDE $p_{A^{(k)},H^{(k)}} \in \mathcal{A}$ for the next iteration. To perform APPEX on a general additive noise SDE (2), the spaces $\mathcal{A}$ and the drift maximum likelihood parameter estimation step (27) would be modified accordingly. We show in Corollary 27 of the appendix that there is a closed form diffusion MLE for general additive noise SDEs, once the drift has been estimated.

Previous works implement similar iterative schemes (Zhang, 2024; Shen et al., 2025; Vargas et al., 2021), which alternate between trajectory inference and MLE parameter estimation. In contrast to our method, each of these works assumes that the diffusion of the process is known, and hence invariant across iterations. One reason for this assumption is to ensure finite KL divergence between the estimated trajectories and estimated SDEs for all iterations (Vargas et al., 2021). Given continuously observed marginals of a d-dimensional process, the KL divergence between two laws on paths q, p is taken over the path space $\Omega = C([0,T], \mathbb{R}^d)$, such that

$$D_{KL}(q\|p) = \int_\Omega \log\left(\frac{\mathrm{d}q}{\mathrm{d}p}(\omega)\right) \mathrm{d}q(\omega).$$

Thus, $D_{KL}(q\|p)$ will only be finite between two laws on paths if we can define the Radon-Nikodym derivative $\frac{\mathrm{d}q}{\mathrm{d}p}$ over path space Ω. By Girsanov's theorem, this is only ensured when both processes share the same diffusion (Shen et al., 2025; Vargas et al., 2021). However, if we only consider measurements from a finite number of observed marginals, the KL divergence of the discretized processes q^N, p^N over the path space projected on the N measurement times, $\Sigma_{\mathcal{D}} = C(\{t_i\}_{i=0}^{N-1}, \mathbb{R}^d)$, can be decomposed as follows, see Benamou et al. (2019):

$$D_{KL}(q^N\|p^N) = \sum_{i=0}^{N-2} D_{KL}(q_{i,i+1}\|p_{i,i+1}) - \sum_{i=1}^{N-2} D_{KL}(q_i\|p_i), \tag{29}$$

where $p_{i,i+1}, q_{i,i+1}$ are the joint probability measures of q, p, restricted to times t_i and t_{i+1}, and similarly q_i, p_i are the marginals of q, p at time t_i. We note that the above equality holds since both p^N and q^N are Markovian. Then, since $D_{KL}(q_i\|p_i) \leq D_{KL}(q_{i,i+1}\|p_{i,i+1})$ holds by the data-processing inequality, it follows that $D_{KL}(q^N\|p^N)$ will be finite as long as $D_{KL}(q_{i,i+1}\|p_{i,i+1}) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \log\left(\frac{\mathrm{d}q_{i,i+1}}{\mathrm{d}p_{i,i+1}}(x)\right) q_{i,i+1}(x)\,\mathrm{d}x < \infty$ for each $i = 0, \ldots, N-2$. In particular, if p and q are the laws of two different drift-diffusion SDEs with distinct non-degenerate diffusions, then the KL divergence $D_{KL}(q^N\|p^N)$ over the discretized path space $\Sigma_{\mathcal{D}}$ will be finite. Since we consider a setting with a finite number of temporal marginal observations in this work, we may consider laws on paths discretized over these observations. This allows us to both consider diffusions which are not known in advance, and to improve our diffusion estimates with respect to KL divergence after each iteration.

5.1 Trajectory Inference Via Anisotropic Entropy-Regularized Optimal Transport

To begin each iteration of APPEX, we infer trajectories given the set of temporal marginals $\{p_i\}_{i=0}^{N-1}$ and a reference SDE $p_{A^{(k-1)},H^{(k-1)}} \in \mathcal{A}$, by solving an associated Schrödinger Bridge problem. Under certain conditions, the SB problem (5) can be solved using the variational entropic optimal transport (EOT) formulation

$$\pi^* = \underset{\pi \in \Pi(\mu,\nu)}{\arg\min} \int c(x,y)\, \mathrm{d}\pi(x,y) + \epsilon^2 D_{KL}(\pi \| \mu \otimes \nu). \tag{30}$$

Intuitively, π^* transports the probability measure μ to ν while minimizing a total cost, based on the function $c(x,y)$ and the entropic regularization $\epsilon^2 > 0$. Objectives (5) and (30) are equivalent (Peyré et al., 2019; Zhang, 2024), under the reference measure $d\mathcal{K}(x,y) = K(x,y)d\mu(x)d\nu(y)$ with density $K(x,y) \propto e^{\frac{-c(x,y)}{\epsilon^2}}$. The convention in the literature is to set $c(x,y) = \frac{\|y-x\|^2}{2}$, to define an isotropic Gaussian law $\mathcal{K}$ (Janati et al., 2020; Lavenant et al., 2021; Chizat et al., 2022; Zhang, 2024). The EOT problem (30) can then be solved via Sinkhorn's algorithm.

However, the standard EOT problem (30) only considers a scalar regularization parameter ϵ^2, which means that it can only directly model reference SDEs with isotropic diffusion (see Remark 26 in the appendix). To generalize trajectory inference for SDEs with anisotropic diffusion, we formalize "anisotropic entropy-regularized optimal transport" (AEOT), by considering transition kernels K with custom mean and covariance, e.g. parameterizing K with A and $H = GG^\top$ rather than the scalar $\epsilon^2 > 0$. Indeed, we can consider more general Gaussian transition kernels, $K_\theta(x,y) = \exp(\frac{(y-\mu(x,\theta))^\top \Sigma(\theta)^{-1}(y-\mu(x,\theta))}{-2})$, in order to model a transition $y|x \sim \mathcal{N}(\mu(x,\theta), \Sigma(\theta))$, whose covariance $\Sigma(\theta)$ is possibly anisotropic. Under this formulation, the cost is given by the inner product $c_\theta(x,y) = \frac{(y-\mu(x,\theta))^\top \Sigma(\theta)^{-1}(y-\mu(x,\theta))}{2}$. Trajectory inference would amount to solving the AEOT problem:

$$\pi^* = \underset{\pi \in \Pi(\mu,\nu)}{\arg\min} \int \frac{(y-\mu(x,\theta))^\top \Sigma(\theta)^{-1}(y-\mu(x,\theta))}{2}\, \mathrm{d}\pi(x,y) + D_{KL}(\pi \| \mu \otimes \nu), \tag{31}$$

where the entropic regularization is captured in the matrix $\Sigma(\theta)$ and function $\mu(x,\theta)$. The AEOT problem over a pair of marginals (31) can be solved via Sinkhorn's algorithm with inputs μ, ν, K_θ. Proposition 25 in the Appendix then shows that the solution to the multi-marginal trajectory inference step (26) is the joint distribution given by the Markov concatenation of couplings (see Definition 24 in the Appendix)

$$\pi^* = \pi_{N-2,N-1} \circ \ldots \circ \pi_{0,1},$$

where $\pi_{i,i+1}$ is the AEOT solution (31) with marginals $\mu = p_i, \nu = p_{i+1}$ and a transition kernel determined by the reference SDE.

Remark 10 (Application to empirical marginals) *Proposition 25 is formalized for exact temporal marginals p_i. However, in practice, we observe empirical measures $\hat{p}_i$, which converge in distribution to p_i, in the limit of infinite data. It is for these empirical measures*

$\hat{p}_i$ that we actually compute the estimated couplings $\hat{\pi}_{i,i+1}$ in practice. This corresponds to numerically estimating the AEOT solution

$$\hat{\pi}_{i,i+1} = \underset{\pi \in \Pi(\hat{p}_i, \hat{p}_{i+1})}{\arg\min} \int \frac{(y-\mu(x,\theta))^\top \Sigma(\theta)^{-1}(y-\mu(x,\theta))}{2} \, \mathrm{d}\pi(x,y) + D_{KL}(\pi \| \hat{p}_i \otimes \hat{p}_{i+1}). \tag{32}$$

By the main theorem in Ghosal et al. (2022), it holds that $\hat{\pi}_{i,i+1}$ converges in distribution almost surely to the minimizer of (31) with $\mu = p_i$ and $\nu = p_{i+1}$ as the number of samples goes to infinity for each time i and $i+1$. Combining this with the previous proposition, we see that estimating (32) for each pair of times i and $i+1$ is asymptotically equivalent to the KL minimization step in (26).

We also note that (32) *is practically solved using Sinkhorn's algorithm with empirical marginals $\hat{p}_i, \hat{p}_{i+1}$ and the discretized transition kernel K_θ of the reference SDE. For example, if we have M samples per empirical marginal, then $\hat{p}_i \sim$ Unif$(x_{t_i}^{(j)} : j = 1, \dots M)$ and $\hat{p}_{i+1} \sim$ Unif$(x_{t_{i+1}}^{(j)} : j = 1, \dots M)$ would both be discrete uniform distributions over their samples, and $K_\theta \in \mathbb{R}^{M \times M}$ would be a square matrix, such that entry $K_{\theta_{j,k}}$ is obtained by applying the given transition kernel on the data points $x_{t_i}^{(j)}$ and $x_{t_{i+1}}^{(k)}$.*

5.2 Parameter Estimation Via MLE

To optimize objectives (27) and (28) for each iteration of APPEX, we require maximum likelihood estimators for the SDE parameters, given multiple observed trajectories from $[0, T]$. In the context of iteration k of APPEX, these are the trajectories sampled from the law on paths $q^{(k)}$ obtained from the trajectory inference step. We derive closed-form maximum likelihood estimators for the linear additive noise SDE from Equation 3, given multiple trajectories in Proposition 11. Closed form drift MLE solutions are generally unavailable for additive noise SDEs, however, we derive a closed form diffusion estimate (which depends on the drift) for general additive noise SDEs in Corollary 27 of the appendix.

Proposition 11 (MLE estimators for drift and diffusion from trajectories) *Given M trajectories over N equally spaced times: $\{X_{i\Delta t}^{(j)} : i \in 0, ..., N-1,\ j \in 0, ..., M-1\}$ sampled from the linear additive noise SDE* (3)*, the maximum likelihood solution for linear drift is approximated by*

$$\hat{A} = \frac{1}{\Delta t} \left(\sum_{i=0}^{N-2} \sum_{j=0}^{M-1} (\Delta X_i^{(j)}) X_i^{(j)^\top} \right) \left(\sum_{i=0}^{N-2} \sum_{j=0}^{M-1} X_i^{(j)} X_i^{(j)^\top} \right)^{-1} \tag{33}$$

$$\overset{M \to \infty}{\longrightarrow} \frac{1}{\Delta t} \left(\sum_{i=0}^{N-2} \mathbb{E}_{p_{i,i+1}}[(\Delta X_i) X_i^\top] \right) \left(\sum_{i=0}^{N-2} \mathbb{E}_{p_i}[X_i X_i^\top] \right)^{-1}, \tag{34}$$

where $p_{i,i+1}$ is the joint measure over X_{t_i} and $X_{t_{i+1}}$. Similarly, the maximum likelihood solution for diffusion is approximated by

$$\hat{H} = \frac{1}{MT} \sum_{i=0}^{N-2} \sum_{j=0}^{M-1} \left((\Delta X_i^{(j)} - AX_i^{(j)} \Delta t)(\Delta X_i^{(j)} - AX_i^{(j)} \Delta t)^\top \right) \tag{35}$$

$$\overset{M\to\infty}{\longrightarrow} \frac{1}{T} \sum_{i=0}^{N-2} \mathbb{E}_{p_{i,i+1}}[(\Delta X_i - AX_i \Delta t)\,(\Delta X_i - AX_i \Delta t)^\top]. \tag{36}$$

Proof See Section F of the Appendix. We note that estimators $\hat{A}$ and $\hat{H}$ were derived using the discretized transition kernel, $X_{i+1}|X_i \sim \mathcal{N}(X_i + AX_i\Delta t, H\Delta t)$. ■

We note that the MLE drift estimator for A does not depend on the diffusion H, but the MLE estimator for H depends on A. We therefore estimate drift first in each iteration of APPEX.

5.3 The APPEX Algorithm

We summarize APPEX's implementation in Algorithm 1. In each iteration, trajectory inference (26) is performed by solving the AEOT problem (31), and the drift and diffusion estimates are updated via MLE. By default, we use Sinkhorn's algorithm to solve AEOT and we use closed form MLE solutions when available.

Algorithm 1 Parameter estimation for an additive noise SDE from temporal marginals with APPEX

Input: Observed marginals $\hat{p}_i$, $i = 0, \ldots, N-1$, number of iterations K, Δt
Result: Estimated drift function $\hat{b}$ and additive noise $\hat{H}$
Initialize: $\hat{b} \leftarrow 0$, $\hat{H} \leftarrow \sigma^2 \mathrm{I}_d$, $k \leftarrow 0$
while $k < K$ **do**
    for $i = 0, \ldots, N-2$ **do**
        $\hat{\pi}_{i,i+1} \leftarrow$ Anisotropic-Entropy-Regularized-Optimal-Transport$(\hat{b}, \hat{H}, \hat{p}_i, \hat{p}_{i+1}, \Delta t)$
    end for
    Sample-Trajectories $\leftarrow \hat{\pi}_{N-2,N-1} \circ \ldots \circ \hat{\pi}_{0,1}(\hat{p}_0)$
    $\hat{b} \leftarrow$ MLEfit(Sample-Trajectories)
    $\hat{H} \leftarrow$ MLEfit(Sample-Trajectories, $\hat{b}$)
    $k \leftarrow k + 1$
end while
$\mathcal{G} \leftarrow$ Estimate-Causal-Graph$(\hat{b}, \hat{H}, \epsilon)$

Remark 12 (Application to causal discovery) *By Lemma 9, we can estimate the causal graph from the SDE parameters learned by APPEX. For nonlinear drift $b(X_t)$, we would include the edge $i \to j$ if the jth component of the drift, b_j, is a function of $X^{(i)}$. This can for example be assessed by analyzing the Jacobian matrix of b. For linear drift, this*

simplifies to analyzing each entry $A_{j,i}$ and choosing a threshold to determine the presence of directed edges. For edges related to latent confounders (from shared diffusion), one can choose a threshold for $|H_{i,j}|$ to determine whether there is an unobserved confounder causing $X^{(i)}$ and $X^{(j)}$. While thresholding is one of the simplest methods to extract a causal graph from learned SDE parameters, we note that many other methods can be used for this task, and thus leave the implementation as general as possible.

Because the trajectory inference and parameter estimation subprocedures are optimal with respect to minimizing KL divergence, we can show that APPEX's estimates progressively fit the data better with each iteration. Furthermore, this data-fitting objective is uniquely minimized by the true parameters (A, H), if identifiability conditions hold.

Lemma 13 *Suppose that A, H are the true parameters of a linear additive noise SDE with temporal marginals $p_0, ..., p_{N-1}$. Let $q^{(k)} \in \mathcal{D}$ be the law of the estimated trajectories at iteration k of APPEX, and $p_{A^k,H^k} \in \mathcal{A}$ be the law of the estimated linear additive noise SDE. Then, the relative entropy between the estimated law of the trajectories and the law of the estimated linear additive noise SDE is decreasing with each iteration (Shen et al., 2024):*

$$D_{KL}(q^{(k+1)} \| p_{A^{k+1},H^{k+1}}) \leq D_{KL}(q^{(k)} \| p_{A^k,H^k}) \quad \forall k \geq 0.$$

Furthermore, if p_0 is not weakly auto-rotationally invariant, then as we let the number of observed marginals $N \to \infty$ for any sampling scheme that would reduce the maximum time gap between marginals to 0, the relative entropy between the estimated law of the trajectories and the law of the estimated SDE is uniquely minimized by the underlying SDE $p_{A,H}$:

$$\inf_{q \in \mathcal{D}, p \in \mathcal{A}} D_{KL}(q \| p) = 0 \iff q = p = p_{A,H}.$$

Proof Proposition 25 shows that $q^{(k+1)}$ minimizes relative entropy to p_{A^k,H^k}. Similarly, any MLE solution $p_{A^{(k+1)},H^{(k+1)}}$ minimizes relative entropy to $q^{(k+1)}$. Proposition 11 approximates the MLE parameters A^{k+1} and $H^{(k+1)}$. By construction of the iterative scheme,

$$\begin{aligned} D_{KL}\left(q^{(k+1)} \,\middle\|\, p_{A^{(k+1)},H^{(k+1)}}\right) &\leq D_{KL}\left(q^{(k+1)} \,\middle\|\, p_{A^{(k+1)},H^{(k)}}\right) \\ &\leq D_{KL}\left(q^{(k+1)} \,\middle\|\, p_{A^{(k)},H^{(k)}}\right) \leq D_{KL}\left(q^{(k)} \,\middle\|\, p_{A^{(k)},H^{(k)}}\right), \end{aligned}$$

which proves the first claim of the lemma. If X_0 is not weakly auto-rotationally invariant, then $p_{A,H}$ is the law of the unique linear additive noise SDE, which obeys the marginal constraints $p_0, \dots p_{N-1}$ as $N \to \infty$ (Theorem 5). Equivalently, we have $\mathcal{A} \cap \mathcal{D} = \{p_{A,H}\}$,

$$\inf_{q \in \mathcal{D}, p \in \mathcal{A}} D_{KL}(q \| p) = 0 \iff q = p = p_{A,H}.$$

■

6. Experiments

In a wide range of experiments on simulated data from linear additive noise SDEs, we find that APPEX accurately identifies the underlying SDE from its temporal marginals and that it outperforms the popular Waddington-OT method (Schiebinger et al., 2019). In Section 6.1, we assess the validity of Theorem 5 by showing that APPEX can learn the drift and diffusion parameters of classical non-identifiable SDEs, if X_0 is not auto-rotationally invariant. We then demonstrate APPEX's effectiveness on randomly generated higher-dimensional SDEs in Section 6.2. Finally, in Section 6.3, we show that APPEX can use its parameter estimates to recover the underlying causal graph, thus supporting Lemma 9. An empirical experiment for consistency is also provided in Appendix G, which shows that as the number of samples per time marginal increases, APPEX's drift and diffusion estimates converge to the true parameters.

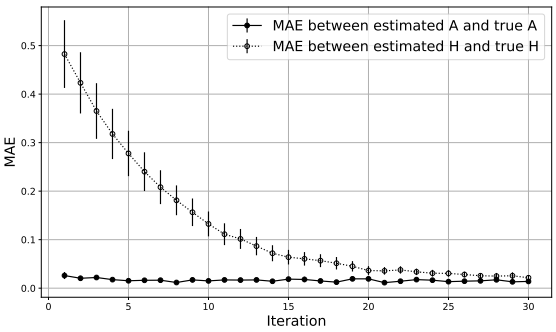
For each experiment, we simulate data for linear additive noise SDEs using Euler-Maruyama discretization with $dt_{EM} = 0.01$, such that we generate $M = 500$ trajectories, each observed across 100 time steps. Each d-dimensional trajectory is initialized with $p_0 \sim \text{Unif}\{x_i\}_{i=1}^d$ where $\{x_i\}_{i=1}^d$ are random linearly independent vectors with each entry having a magnitude uniformly sampled between 2 and 10. Since these are discrete random variables whose support spans the full space, it follows from Proposition 20 and Theorem 5 that the drift and diffusion are theoretically identifiable from marginals. To model the "marginals-only" setting, we subsample at the rate $dt = 0.05$ to produce $N = 20$ marginals with $M = 500$ observations per time.

To perform parameter estimation, we use 30 iterations of our APPEX algorithm, such that the first reference SDE is an isotropic Brownian motion $\mathrm{d}X_t = \sigma \mathrm{d}W_t$, even when the true process has anisotropic noise. To model the realistic setting where diffusion is not precisely known, we assume that the initial guess of the diffusion's trace is within an order of magnitude of the ground truth trace $\text{tr}(H)$. Specifically, we initialize $A^{(0)} = 0, H^{(0)} = \sigma^2 \mathrm{I}_d \; s.t. \; \sigma^2 \sim (\text{tr}(H)/d)10^{\text{Unif}(-1,1)}$. Because the first reference SDE is $\sigma \mathrm{d}W_t$, we note that the first iteration of APPEX is equivalent to performing MLE on trajectories that were inferred using the Waddington-OT (WOT) method from Schiebinger et al. (2019), based on standard entropy regularized OT (30) with $\epsilon^2 = \sigma^2 \Delta t$. APPEX and WOT are distinguished by the fact that APPEX benefits from further iterations, and allows the reference SDE to have anisotropic diffusion and non-zero drift. For each iteration of APPEX, we use Sinkhorn's algorithm to solve the anisotropic entropy-regularized optimal transport problem for trajectory inference, and we use the closed form MLE solutions from Proposition 11. Due to time complexity and numerical stability, we use linearized discretizations for the Gaussian transition kernels. The Python code for reproducing the experimental results and figures is available at `https://github.com/guanton/APPEX`. Analogous code in R is available at `https://github.com/HydroML/X0isAllYouNeed`.

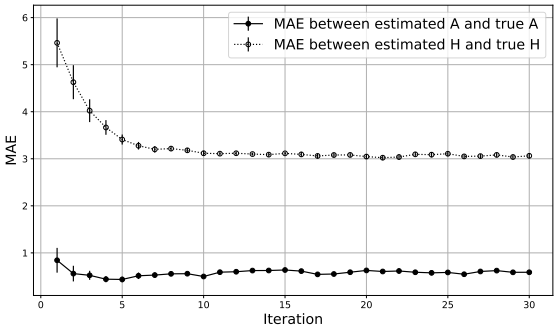
6.1 Revisiting Non-Identifiable SDEs

We consider three non-identifiable pairs of SDEs from the literature (summarized in Appendix B), but revise X_0 to be non-auto-rotationally invariant, such that by Theorem 5, the SDE parameters are now theoretically identifiable from observed marginals. 10 replicates of each experiment were performed, such that the replicates featured data from different

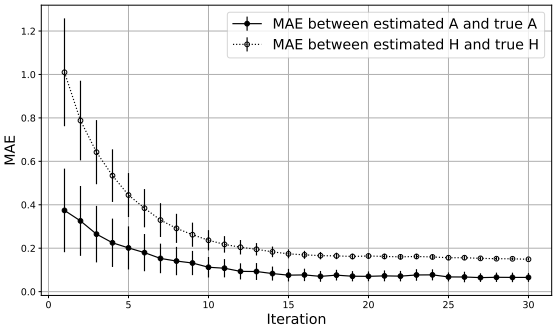
valid p_0 initializations and different initial diffusivities σ^2. We track the mean absolute error (MAE), plotted in Figure 4, between the true drift/diffusion parameters and their estimates.



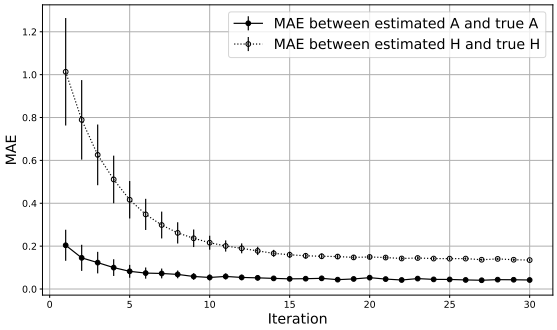
(a) $\mathrm{d}X_t = -X_t\,\mathrm{d}t + \mathrm{d}W_t$ from Example 1



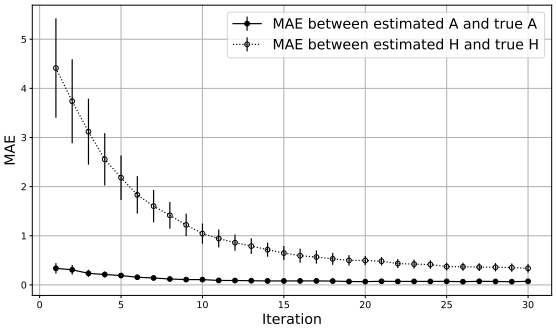
(b) $\mathrm{d}X_t = -10X_t\,\mathrm{d}t + \sqrt{10}\mathrm{d}W_t$ from Ex. 1



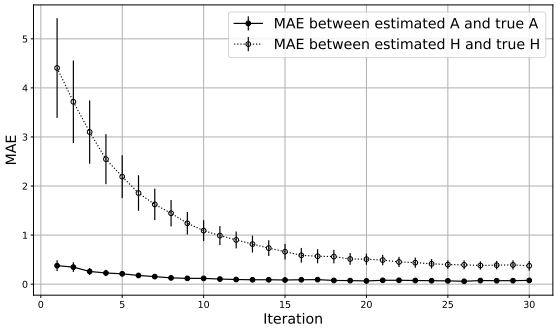
(c) $\mathrm{d}X_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \mathrm{d}W_t$ from Ex. 2



(d) $\mathrm{d}X_t = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} X_t\mathrm{d}t + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \mathrm{d}W_t$ from Ex. 2



(e) $\mathrm{d}X_t = \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix} X_t\mathrm{d}t + \begin{bmatrix} 1 & 2 \\ -1 & -2 \end{bmatrix} \mathrm{d}W_t$ from Ex.3



(f) $\mathrm{d}X_t = \begin{bmatrix} \frac{1}{3} & \frac{4}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix} X_t\mathrm{d}t + \begin{bmatrix} 1 & 2 \\ -1 & -2 \end{bmatrix} \mathrm{d}W_t$ from Ex.3

Figure 4: The mean absolute error for estimates of A and H using APPEX is shown per iteration for all three pairs of SDEs from Section B.

The results are consistent with Theorem 5 and Lemma 13. In each example, APPEX is able to estimate the drift and diffusion by iteratively improving upon both estimates, as evidenced by decreasing MAEs as the number of iterations increases. In particular, we observe the worst performance from the first iteration, which is equivalent to performing parameter estimation following trajectory inference by WOT. The most difficult SDE to infer was $\mathrm{d}X_t = -10X_t\,\mathrm{d}t + \sqrt{10}\mathrm{d}W_t$, due to the high noise scale. We also note that APPEX succeeded in estimating the degenerate diffusion from the SDEs in Example 3, despite the fact that degenerate diffusion can result in infinite KL divergence with respect to a misspecified reference SDE.

We emphasize that if we observe marginals such that the SDE is non-identifiable, for example, inferring the Ornstein-Uhlenbeck SDEs $\mathrm{d}X_t = -X_t\,\mathrm{d}t + \mathrm{d}W_t$ from snapshots sampled from its equilibrium distribution $\mathcal{N}(0, \frac{1}{2})$, then no method can recover the true parameters. Indeed, in this example, an infinite number of linear additive noise SDEs share the same equilibrium distribution (as long as the drift to diffusion ratio is 1), and are hence consistent with the observed data. When using APPEX to estimate the SDE in this non-identifiable setting, we indeed saw that it recovered different drift and diffusion estimates, with ratio close to 1 (see Figure 8).

6.2 Higher Dimensional Randomly Generated Linear Additive Noise SDEs

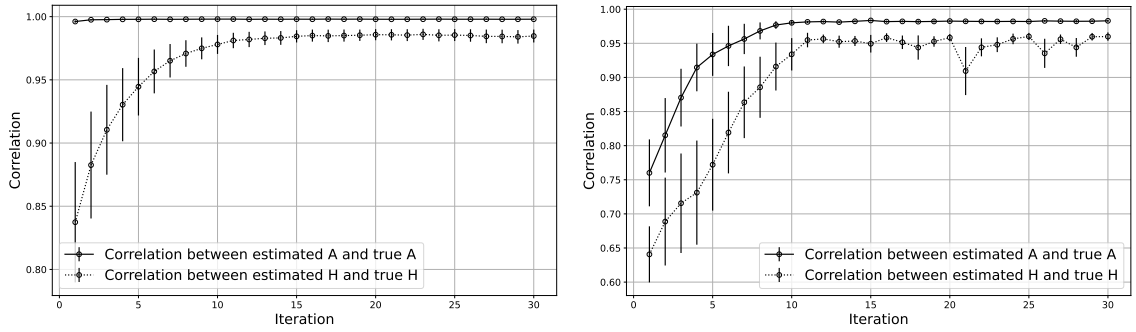
We now consider parameter estimation on a broad range of higher dimensional random linear additive noise SDEs. For each dimension $d = 3, \ldots, 10$, we generate 10 random SDEs. To create each drift matrix A, we randomly sample each of its d^2 entries from $\mathrm{Unif}(-5, 5)$, and verify that its principal eigenvalue is less than 1. This allows us to consider non-stationary processes, while obeying reasonable growth conditions. To create each diffusion G, we randomly initialize each of its d^2 entries via $\mathrm{Unif}(-1, 1)$ and then set $H = GG^\top$. For numerical stability, we perform Sinkhorn on the logarithmic scale for this experiment. We evaluate each estimate's performance using MAE and correlation with the true parameters, where correlation is computed by vectorizing the corresponding matrices. We report the mean and standard error of the MAE across the 10 runs.

d	**A Estimation**		**$GG^\top$ Estimation**	
	WOT	**APPEX**	**WOT**	**APPEX**
3	0.351 ± 0.04	$\mathbf{0.237 \pm 0.04}$	0.793 ± 0.205	$\mathbf{0.147 \pm 0.030}$
4	0.730 ± 0.067	$\mathbf{0.328 \pm 0.041}$	1.549 ± 0.439	$\mathbf{0.415 \pm 0.070}$
5	0.912 ± 0.060	$\mathbf{0.602 \pm 0.195}$	2.174 ± 0.702	$\mathbf{0.362 \pm 0.039}$
6	1.43 ± 0.170	$\mathbf{0.358 \pm 0.020}$	18.010 ± 8.636	$\mathbf{0.256 \pm 0.046}$
7	1.480 ± 0.132	$\mathbf{0.360 \pm 0.015}$	6.807 ± 1.724	$\mathbf{0.345 \pm 0.037}$
8	1.862 ± 0.137	$\mathbf{0.460 \pm 0.015}$	5.472 ± 1.266	$\mathbf{0.359 \pm 0.019}$
9	1.803 ± 0.222	$\mathbf{0.487 \pm 0.016}$	8.134 ± 3.024	$\mathbf{0.454 \pm 0.122}$
10	1.670 ± 0.241	$\mathbf{0.439 \pm 0.019}$	35.122 ± 28.529	$\mathbf{0.317 \pm 0.025}$

Table 1: Mean absolute error (MAE) of estimated drift and diffusion for dimensions d. Each box reports the mean $\pm$ the standard error.

d	**A Estimation**		**$GG^\top$ Estimation**	
	WOT	**APPEX**	**WOT**	**APPEX**
3	0.996 ± 0.001	$\mathbf{0.998 \pm 0.001}$	0.837 ± 0.048	$\mathbf{0.985 \pm 0.005}$
4	0.943 ± 0.015	$\mathbf{0.987 \pm 0.005}$	0.729 ± 0.039	$\mathbf{0.865 \pm 0.031}$
5	0.921 ± 0.016	$\mathbf{0.952 \pm 0.030}$	0.728 ± 0.040	$\mathbf{0.909 \pm 0.018}$
6	0.794 ± 0.040	$\mathbf{0.986 \pm 0.001}$	0.530 ± 0.056	$\mathbf{0.961 \pm 0.007}$
7	0.792 ± 0.029	$\mathbf{0.988 \pm 0.001}$	0.595 ± 0.037	$\mathbf{0.946 \pm 0.012}$
8	0.699 ± 0.035	$\mathbf{0.981 \pm 0.002}$	0.611 ± 0.042	$\mathbf{0.949 \pm 0.006}$
9	0.740 ± 0.033	$\mathbf{0.978 \pm 0.002}$	0.615 ± 0.025	$\mathbf{0.919 \pm 0.033}$
10	0.760 ± 0.049	$\mathbf{0.983 \pm 0.001}$	0.641 ± 0.041	$\mathbf{0.960 \pm 0.006}$

Table 2: Correlation between estimated and true drift and diffusion for dimensions d. Each box reports the mean $\pm$ the standard error.



(a) $d = 3$ (b) $d = 10$

Figure 5: The correlation between the estimated and true SDE parameters is plotted per iteration across 10 random linear additive noise SDEs for dimensions 3 and 10. Each point reports the mean of the 10 runs, while the bars represent the standard error.

The results, given in Tables 1 and 2, demonstrate that APPEX estimates both SDE parameter sets robustly across all settings. Importantly, APPEX can handle arbitrary additive noise structures $H = GG^\top$, as evidenced by the fact that its drift and diffusion estimates have low MAE and high correlation with respect to the true parameters. This is significant because previous literature has focused on the setting of isotropic noise, but in practice, we expect correlated noise structures, as well as unequal noise along the main diagonal. In contrast, although WOT estimates the drift somewhat decently, it is unable to estimate the diffusion accurately, particularly in higher dimensions, since it is constrained to isotropic noise in its reference SDE. Figure 5 shows how APPEX is able to re-orient incorrect diffusion priors to closely match the true diffusion of the underlying SDE. While running increasing numbers of iterations within APPEX generally leads to monotonic increases in parameter accuracy (as measured by correlation in Figure 5), we observe occasional instability due to the combination of finite observations, numerical limitations during trajectory inference from capping the maximum number of iterations of Sinkhorn, and the fact that the algorithm is building on previous (and wrong) estimates.

6.3 Causal Discovery

We conclude with experiments that test our method's ability to recover the underlying causal graph. We first consider causal discovery from systems without latent confounders, and then consider causal discovery from systems with latent pairwise confounders through correlated diffusion.

In both settings, we generate random SDEs of dimension $d = 3, 5, 10$ whose causal structure is determined by Erdos-Renyi graphs $\mathcal{G}(d, p)$, such that each of the $d(d-1)$ possible directed edges are included with probability p. As in Section 6.2, we ensure that the maximal eigenvalue of the randomly generated ground truth drift matrix is at most 1 to prevent unbounded growth. Lemma 9 states that directed edges $e = i \to j \in E$ in the causal graph $\mathcal{G} = ([d], E, \tilde{E})$ are characterized by the condition $A_{j,i} \neq 0$. To simulate ground truth edges, we follow standard convention, by simulating edge weights $A_{j,i}$ uniformly via $\text{Unif}(-5, -0.5) \cup \text{Unif}(0.5, 5)$ (Runge, 2021; Reisach et al., 2021). We infer the presence of an edge $i \to j$, if in the estimated drift, $\hat{A}_{j,i} > \epsilon$ (corresponding to a positive edge weight) or $\hat{A}_{j,i} < -\epsilon$ (corresponding to a negative edge weight). We choose our threshold $\epsilon = 0.5$ to match the minimum edge weight magnitude $|A_{j,i}| \geq 0.5$ in our simulations. Inference accuracy of directed edges is then measured according to the structural Hamming distance between $\mathcal{G}$ and the estimated graph $\hat{\mathcal{G}}$, which considers the classification of positive directed edges, negative directed edges, and absent edges. The structural Hamming distance adds 1 for every misclassified edge:

$$d(\mathcal{G}, \hat{\mathcal{G}}) = \sum_{(i,j) \in [d] \times [d]} \mathbf{1}\left\{ \text{sgn}(A_{j,i}) \neq \text{sgn}(\hat{A}_{j,i}) \mathbf{1}_{|\hat{A}_{j,i}| > \epsilon} \right\}, \tag{37}$$

where the sum ranges over all ordered pairs including $i = j$, so that self-regulatory edges $i \to i$ are also included. If the observed system does not have any latent confounders from correlated diffusion, we enforce this by setting the diffusion matrix G to be strictly diagonal. This ensures that the only edges in $\mathcal{G}$ are uni-directional edges from the drift. We set each of the d diagonal entries of G via $\text{Unif}(0.1, 1)$. Conversely, when modeling pairwise latent confounders, we introduce correlated diffusion by selecting a random subset of columns in G to have precisely two nonzero entries. In particular, we first initialize all entries along the main diagonal of G to be 1, and for each of the selected columns, we set one non-diagonal entry $G_{i,j} = 1$, to model the pairwise latent confounder $U_{i,j}$. All other entries are set to 0. The number of selected columns with pairwise latent confounders is sampled from $\text{Unif}\{1, \cdots \lfloor 2d/3 \rfloor\}$, e.g. we sample from $\text{Unif}\{1, 2\}$ for $d = 3$, $\text{Unif}\{1, 2, 3\}$ for $d = 5$, and $\text{Unif}\{1, \ldots 6\}$ for $d = 10$. Lemma 9 states that the condition $H_{i,j} = G_{i,\cdot} G_{j,\cdot}^\top \neq 0$ identifies pairwise latent confounders $U_{i,j}$. Since all entries of G are nonnegative in this construction, no cancellation occurs in $H = GG^\top$, and the confounded pairs are exactly those with $H_{i,j} \neq 0$, for which $H_{i,j} \geq 1$. We read the ground-truth adjacency directly off the construction, and infer the presence of $U_{i,j}$ if the estimated diffusion satisfies $|\hat{H}_{i,j}| > \epsilon_H$, with $\epsilon_H = 0.5$. Inference of edges from pairwise latent confounders is measured according to the structural Hamming distance, which adds 1 for every misclassified pairwise latent confounder:

$$d_{\text{latent}}(\mathcal{G}, \hat{\mathcal{G}}) = \sum_{1 \leq i < j \leq d} \mathbf{1}\left\{ \mathbf{1}\{H_{ij} \neq 0\} \neq \mathbf{1}\{|\hat{H}_{ij}| > \epsilon_H\} \right\}. \tag{38}$$

Table 3 reports the mean structural Hamming distances for directed edges from causal graphs estimated by WOT and APPEX, in the setting without latent confounders. APPEX outperforms WOT across all tested dimensions $d \in \{3, 5, 10\}$ and edge probabilities $p \in \{0.1, 0.25, 0.5\}$, with the discrepancy becoming more pronounced as the causal structure grows in complexity at higher dimensions and edge densities. We observe that APPEX can consistently infer complex causal relationships, whereas WOT's inference is often limited, due to misspecified diffusion. For example, the first row in Figure 6 demonstrates that APPEX recovers the system's negative feedback structure, while WOT introduces additional cycles in order to compensate for a poorly specified reference diffusion.

Table 4 reports the mean structural Hamming distances for drift-related edges and edges from pairwise latent confounders from causal graphs estimated by WOT and APPEX, in the setting with correlated diffusion. The probability p of connecting a directed edge $i \to j$ was fixed at $p = 0.25$. As in the setting with uncorrelated diffusion, APPEX outperforms WOT in all evaluated settings, particularly for higher-dimensional systems.

d	$p = 0.1$		$p = 0.25$		$p = 0.5$	
	WOT	**APPEX**	**WOT**	**APPEX**	**WOT**	**APPEX**
3	0.40 ± 0.40	$\mathbf{0.00 \pm 0.00}$	1.80 ± 0.76	$\mathbf{0.30 \pm 0.30}$	1.40 ± 0.65	$\mathbf{0.00 \pm 0.00}$
5	3.80 ± 2.13	$\mathbf{0.10 \pm 0.10}$	3.56 ± 1.70	$\mathbf{0.00 \pm 0.00}$	4.40 ± 1.14	$\mathbf{0.60 \pm 0.22}$
10	38.3 ± 8.53	$\mathbf{0.50 \pm 0.40}$	48.6 ± 6.55	$\mathbf{0.80 \pm 0.39}$	37.4 ± 4.66	$\mathbf{3.30 \pm 0.70}$

Table 3: Average Structural Hamming Distance (lower is better) with varying dimensions d and random edge probabilities p. Each box reports the mean $\pm$ the standard error.

7. Discussion

Many recent works have focused on inferring the drift or diffusion from an SDE's temporal marginals, yet the problem of non-identifiability remained largely unaddressed (Weinreb et al., 2018; Lavenant et al., 2021; Chizat et al., 2022; Shen et al., 2024; Zhang, 2024; Terpin et al., 2024; Brogat-Motte et al., 2024). To resolve this issue, we first provide necessary and sufficient identifiability conditions for linear additive noise SDEs in Theorem 5, which shows that non-identifiability of the drift and diffusion parameters is feasible if and only if the initial distribution X_0 is weakly auto-rotationally invariant. Building off Theorem 5, we introduce Theorem 6, which proves that the set of drift and diffusion parameters that

d	Hamming distance for drift edges		Hamming distance for latent confounders	
	WOT	**APPEX**	**WOT**	**APPEX**
3	1.20 ± 0.51	$\mathbf{0.7 \pm 0.40}$	1.00 ± 0.30	$\mathbf{0.00 \pm 0.00}$
5	4.80 ± 1.47	$\mathbf{1.60 \pm 0.87}$	3.50 ± 0.64	$\mathbf{0.10 \pm 0.10}$
10	38.00 ± 6.30	$\mathbf{3.00 \pm 0.77}$	23.80 ± 3.86	$\mathbf{2.90 \pm 1.39}$

Table 4: Average Structural Hamming Distance for directed edges, and latent confounders (lower is better) with varying dimensions d and random edge probability $p = 0.25$. Each box reports the mean $\pm$ the standard error.

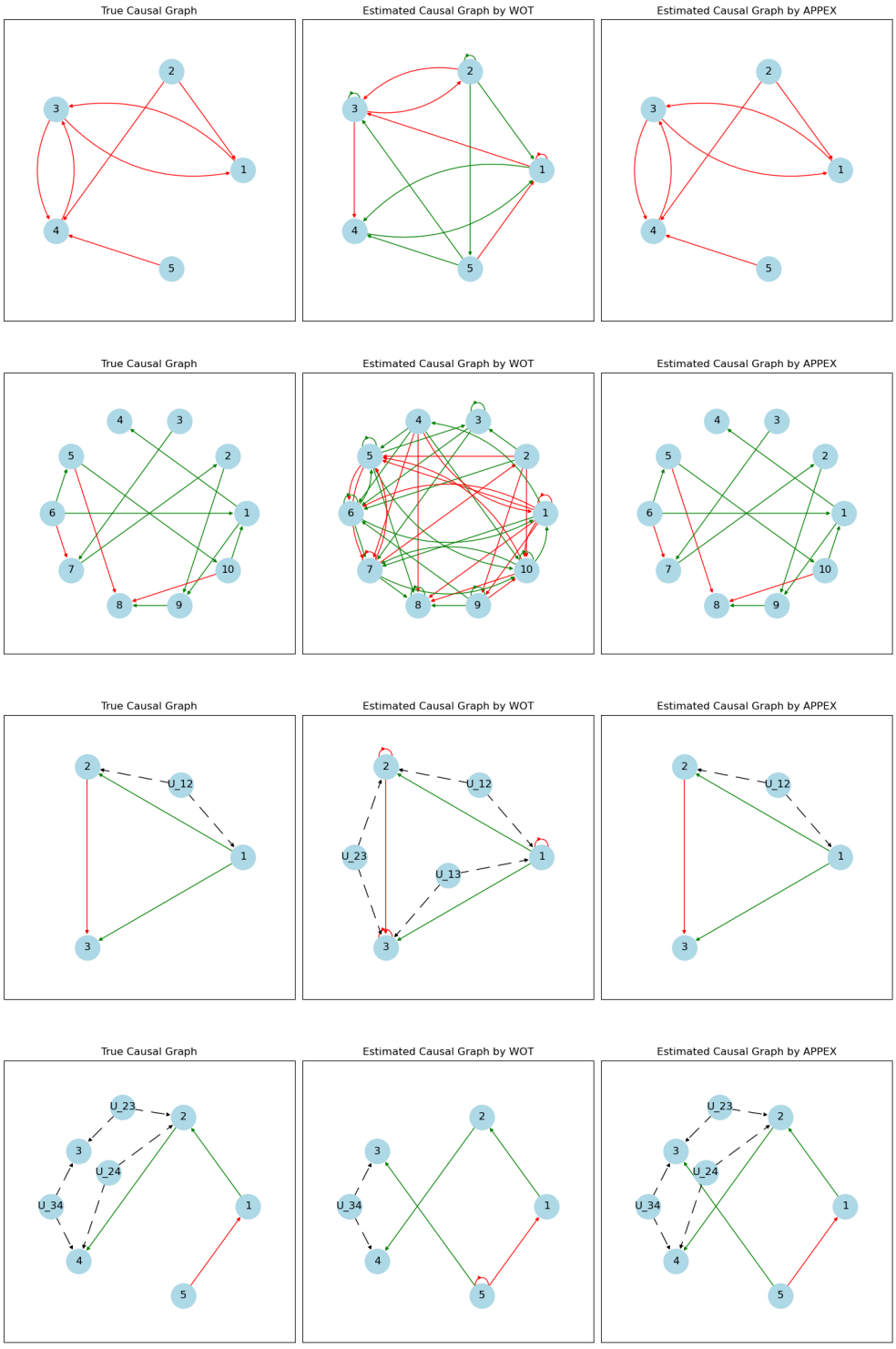


Figure 6: The true and estimated causal graphs by WOT and APPEX for four random SDEs of varying dimensions are illustrated. Red edges $i \to j$ represent negative edge weights, such that $X^{(i)}$ negatively regulates $X^{(j)}$. Green edges $i \to j$ represent positive edge weights, such that $X^{(i)}$ positively regulates $X^{(j)}$. Dashed edges represent the effect of a latent confounder $U_{i,j}$ on observed variables i, j through correlated diffusion.

produce non-identifiable observations has measure zero, even if the initial distribution obeys auto-rotational invariance. Our theory therefore clarifies that non-identifiability of linear drift and diffusion from the process's marginals can only be produced under exceedingly rare circumstances. We complement our identifiability theory by showing that directed edges and latent confounders in the system's causal graph can be recovered from the drift and diffusion parameters, for both linear and nonlinear additive noise SDEs. This provides a theoretical foundation for learning the precise dynamics and causal structure of a linear additive noise process, solely from its temporal marginals, without relying on the usual prior knowledge assumptions and model restrictions.

To build on our theoretical contributions, we also introduce APPEX, a method that simultaneously learns drift and diffusion from temporal marginals. APPEX iterates towards the optimal parameter set by alternating between trajectory inference, via anisotropic entropy-regularized optimal transport, and MLE parameter estimation. Since each of these subprocedures are asymptotically optimal (Lemma 13), given a finite set of the process's true marginals, APPEX improves its estimates with each iteration. When tested on hundreds of randomly generated linear additive noise SDEs between dimensions 3 to 10, APPEX is able to effectively recover the process's drift, diffusion, and causal graph, while significantly outperforming Waddington-OT, which is one of the most popular inference methods in single-cell biology.

Our theoretical and algorithmic contributions offer a number of insights to advance statistical inference methods from temporal marginals. First, our framework removes common model restrictions to expand the set of dynamics that can be inferred from marginals. This includes negative feedback loops and repressilator dynamics, which both arise from rotational drift, and correlated noise, which arises from anisotropic diffusion. These features are common in real-world datasets, and are especially relevant for accurate trajectory and GRN inference from single-cell data (Weinreb et al., 2018; Santos et al., 2024). Second, we demonstrate that joint identification of the unique drift and diffusion parameters is theoretically and practically possible in the "marginals-only" setting. Previous and concurrent works either estimate only one parameter given the other (Weinreb et al., 2018; Lavenant et al., 2021; Chizat et al., 2022; Shen et al., 2024; Zhang, 2024; Forrow, 2024) or estimate both without establishing identifiability (Terpin et al., 2024; Brogat-Motte et al., 2024). We note that even if a user is only interested in one parameter, APPEX offers the ability to estimate it, regardless of how the other parameter is initialized. This is particularly significant given that a misspecified diffusion typically leads to a misidentified drift, and vice-versa. Importantly, the identifiability conditions for linear additive noise SDEs are generic, since non-identifiability requires both a pathological initial distribution, which satisfies (weak) auto-rotational invariance, as well as pathologically aligned SDE parameters, which preserve the symmetry invariance across observed marginals. These theoretical insights provide a principled path for developing new and improved methods that can learn the full SDE purely from observational snapshots.

7.1 Limitations and Future Work

Although this paper presents several breakthroughs for the identification of SDEs from marginal snapshots, we acknowledge several limitations and opportunities for future work.

First, our identifiability result is restricted to time homogeneous linear additive noise SDEs. Although this is a popular model for real-world processes, including for GRN inference (Rohbeck et al., 2024; Zhang, 2024), it would be beneficial to characterize identifiability conditions under more general models. We note that characterizing identifiability is particularly challenging for non-additive noise models, where the noise depends on the process X_t, since the same marginal observations can be explained by a nonlinear additive noise SDE and a linear multiplicative noise SDE (Coomer et al., 2022). For this reason, it may be easier to consider identifiability for parametric families of nonlinear drift under the additive noise model. We note that we have already proven results about causal discovery for additive noise SDEs in Section 4 and that APPEX is also formulated for general additive noise SDEs. Although the MLE for nonlinear drift may not have a closed form solution, optimizers like BFGS (Lavenant et al., 2021; Chizat et al., 2022; Wang et al., 2024) and neural networks (Shen et al., 2024) have been successfully used for SDE parameter estimation on real and synthetic data.

Second, our theory considers the infinite data setting, given the process's true marginals $\{p_i\}_{i=0}^{N-1}$, and we have not studied asymptotic convergence rates. In particular, our identifiability result was established given continuously observed marginals, which corresponds to obtaining infinite samples from infinite marginals. Similarly, the theory surrounding the optimality of APPEX assumes that we have infinite samples from a finite set of marginals. We leave the asymptotic analysis of APPEX for future work. Directions for future work include the sample complexity of entropic optimal transport for trajectory inference, which requires asymptotically more samples per marginal as we decrease the entropic regularization (decreasing time granularity in the SDE setting) (Genevay et al., 2019, Theorem 3), and the fact that the MLE of the AR(1) process from multiple trajectories is slightly biased (Ledolter, 2009). We note that APPEX could be combined with previous methods that relax the marginal constraints (see Lavenant et al. (2021)). This could improve estimates in low data settings, particularly in the biological setting where taking measurements destroys samples.

Third, even in the asymptotic setting, the nonconvex nature of APPEX's alternating projection optimization problem makes the prospect of proving convergence difficult. Although our experiments indicate that APPEX's estimates approach the true solution, regardless of the initial guesses for the drift A and observational diffusion H, we have not yet proved convergence due to the non-convexity of the optimization space. A formal proof of convergence and consistency would further strengthen APPEX's theoretical foundation.

Finally, we note that the trajectory inference step of APPEX, which we solve using Sinkhorn's algorithm, is computationally intensive. For M samples per marginal, each iteration of the standard dense Sinkhorn algorithm requires $O(M^2)$ arithmetic operations once the transport kernel has been constructed. Consequently, J iterations require $O(JM^2)$ operations. The required number of iterations depends on the stopping tolerance, the cost matrix, and the entropy regularization parameter (Dvurechensky et al., 2018). We note that accurate trajectory inference relies on a large number of samples (M), a small entropic regularization $\epsilon^2 \sim \delta t$, and potentially high-dimensional data. Our method is therefore impractical on very large datasets, which are prevalent in biological applications. It would therefore be advantageous to solve the associated Schrödinger bridge problem using more efficient methods. We note that recent work makes use of closed form solutions to the SB

problem when X_0 is a degenerate Dirac distribution, to increase computational efficiency (Huang, 2024). We also note that the SB problem can be expressed via forwards and backwards drifts (Vargas et al., 2021), and can therefore be solved via generative diffusion models (De Bortoli et al., 2021; Shi et al., 2024). Thus, when a closed form solution is not known, generative diffusion models may be more efficient solvers for the trajectory inference problem. Concurrent work estimates drift and diffusion using other methods, including JKO energy minimization (Terpin et al., 2024) and reproducing kernel Hilbert spaces to minimize distance to the Fokker-Planck solution (Brogat-Motte et al., 2024). We have not yet compared the accuracy and computational efficiency of these methods against APPEX, and a systematic comparative analysis would be an interesting direction for future research.

Acknowledgments

The authors would like to thank United Therapeutics for supporting this research. GS and AW also acknowledge the support of the Burroughs Wellcome Fund.

Appendix A. Properties of the Linear Additive Noise SDE

Let X_t evolve according to a linear additive noise SDE (3). Then its solution can be obtained using the integrating factor e^{-At} and is given by (Särkkä and Solin, 2019, Sec 4.3):

$$X_t = e^{At}X_0 + \int_0^t e^{A(t-s)}GdW_s. \tag{39}$$

By the independent increments of W_t, we may interpret X_t as being composed of two independent r.v.'s. The first term amounts to a transformation of the initial distribution $X_0 \to e^{At}X_0$ under the noiseless linear ODE. The second term is an Ito integral, which is in fact a mean-0 Gaussian with time-dependent variance $\Sigma_t = \int_0^t e^{A(t-s)}He^{A^\top(t-s)}ds$. Indeed, since the integrand is deterministic, Gaussianity follows from (Oksendal, 2013, Theorem A.7), and the mean and covariance respectively follow from the non-anticipating property and Ito's isometry.

Appendix B. Non-Identifiable Linear Additive Noise SDEs

We summarize three classical examples of non-identifiability from the literature (Lavenant et al., 2021; Shen et al., 2025; Wang et al., 2024; Hashimoto et al., 2016; Weinreb et al., 2018). In all examples, the processes X_t and Y_t share the same time marginals. In Figure 7, we visualize how changing the initial distribution X_0 to be non auto-rotationally invariant resolves the non-identifiability.

Example 1 (Starting at stationary distribution (Lavenant et al., 2021)) *This example is motivated by the fact that multiple SDEs share the same stationary distribution. Hence, if p_0 is initialized at the stationary distribution, then these SDEs will be non-identifiable from their marginals. We provide a simple example.*

$$\mathrm{d}X_t = -X_t\,\mathrm{d}t + \mathrm{d}W_t, \qquad X_0 \sim \mathcal{N}(0, \frac{1}{2}) \tag{40}$$

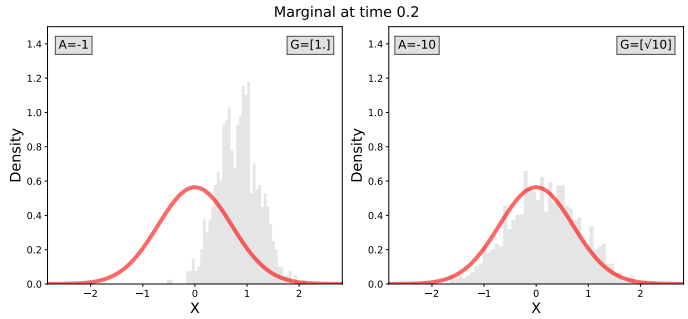
$$\mathrm{d}Y_t = -10Y_t\,\mathrm{d}t + \sqrt{10}\mathrm{d}W_t, \qquad Y_0 \sim \mathcal{N}(0, \frac{1}{2}) \tag{41}$$

Here, both SDEs have the same stationary distribution $\mathcal{N}(0, \frac{1}{2})$ despite having significantly different individual trajectories. Indeed, the stationary distribution of a 1-dimensional 0-mean Ornstein-Uhlenbeck (OU) process with drift $-\lambda X_t$ and diffusion σ is $\mathcal{N}(0, \frac{\sigma^2}{2\lambda})$, which depends only on the drift:diffusivity ratio λ/σ^2 (Lavenant et al., 2021). For multivariate OU processes with drift A and observational diffusion H, the stationary distribution $\mathcal{N}(0, \Sigma)$ depends only on the relationship $\Sigma A^\top + A\Sigma = -H$ (Mandt et al., 2016). We note that this example also generalizes easily to the case where one component of X_0 is a mean-0 Gaussian.

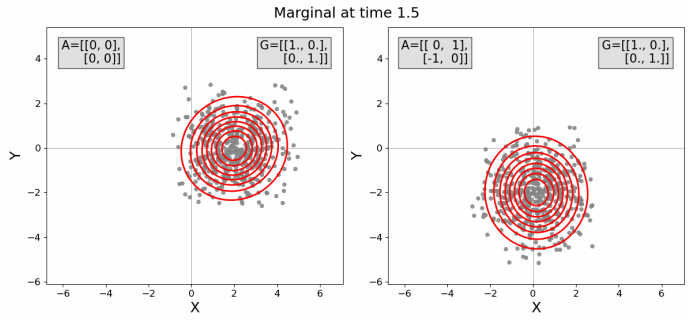
Example 2 (Rotation around process mean (Shen et al., 2025)) *In this example, the first SDE governing X_t is 2-dimensional Brownian motion and the second SDE governing Y_t adds a divergence-free rotational vector field $(x, y) \to (y, -x)$ about $(0, 0)$, which is undetectable if the initial distribution p_0 is rotationally invariant.*

$$\mathrm{d}X_t = \mathrm{d}W_t, \qquad X_0 \sim \mathcal{N}(0, I_d) \tag{42}$$

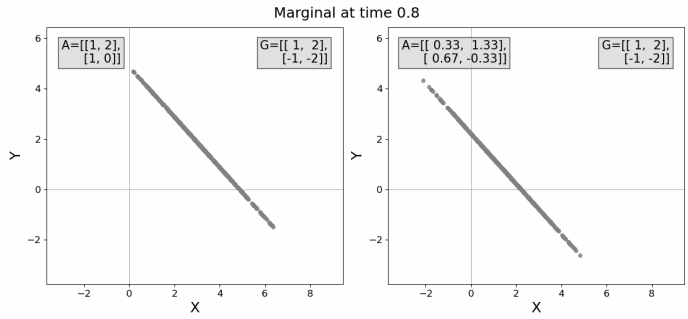
$$\mathrm{d}Y_t = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} Y_t\,\mathrm{d}t + \mathrm{d}W_t, \qquad Y_0 \sim \mathcal{N}(0, I_d) \tag{43}$$



(a) This corresponds to Example 1 with $X_0 \sim \delta_1$. The SDE with $A = -10, H = 10$ (right) has converged to the stationary distribution (red) at time $t = 0.2$, while the SDE with $A = -1, H = 1$ (left) has not yet sufficiently drifted or diffused.



(b) This corresponds to Example 2, with $X_0 \sim \text{Unif}\{(2,0),(2,0.1)\}$. The first SDE is irrotational and maintains mean $(2, 0.05)$. The second SDE (right) rotates about the origin. At time $t = 1.5$, the mean has shifted to approximately $(0, -2)$.



(c) This corresponds to Example 3, with $X_0 \sim \text{Unif}\{(1,0)^\top, (0,1)^\top\}$. The first SDE (left) has eigenvalue 2 for $(2,1)$, while the second (right) has eigenvalue 1, leading to more extreme intercepts in the left plot at $t = 0.8$.

Figure 7: Marginals are plotted at various times for SDE pairs following various non-auto-rotationally invariant initial distributions. Importantly, these initializations enable previously non-identifiable systems to be distinguished.

Hence, if p_0 is an isotropic distribution with mean $(0,0)$, then X_t and Y_t are non-identifiable from one another (Shen et al., 2025). This can also be shown directly with the Fokker-Planck equation (4), since $\nabla \cdot \big((Ax)p(x,t)\big) = \nabla p(x,t) \cdot Ax + p(x,t)\nabla \cdot (Ax) = \nabla p(x,t) \cdot Ax = 0$, since $\nabla \cdot (Ax) = \operatorname{tr}(A) = 0$ and the level sets of p are parallel to the rotational vector field $Ax = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} x$.

Example 3 (Degenerate rank (Wang et al., 2024)) *In this example motivated by Wang et al. (2024), trajectories (and therefore temporal marginals) of distinct SDEs are non-*

identifiable if the process is degenerate, limiting observations to a strict subspace of $\mathbb{R}^d$.

$$\mathrm{d}X_t = \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix} X_t \,\mathrm{d}t + \begin{bmatrix} 1 & 2 \\ -1 & -2 \end{bmatrix} \mathrm{d}W_t, \qquad X_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{44}$$

$$\mathrm{d}Y_t = \begin{bmatrix} 1/3 & 4/3 \\ 2/3 & -1/3 \end{bmatrix} Y_t \,\mathrm{d}t + \begin{bmatrix} 1 & 2 \\ -1 & -2 \end{bmatrix} \mathrm{d}W_t, \qquad Y_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{45}$$

The drift matrices $\begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1/3 & 4/3 \\ 2/3 & -1/3 \end{bmatrix}$ *each have eigenvector* $X_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ *with eigenvalue* -1. *Moreover, the diffusion is rank-degenerate with column space span*$(\begin{bmatrix} 1 \\ -1 \end{bmatrix})$. *Thus, both SDEs will have identical behaviour along span*$(\begin{bmatrix} 1 \\ -1 \end{bmatrix})$ *and are only differentiated by their behaviour elsewhere. Given* $X_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, *the processes will stay within span*$(\begin{bmatrix} 1 \\ -1 \end{bmatrix})$ *and evolve identically. We note that this non-identifiability holds even when we observe trajectories (Wang et al., 2024), whereas the first two examples are identifiable from trajectories but not identifiable from marginals.*

B.1 Properties of Auto-Rotationally Invariant Random Variables

In this section, we present additional theory related to Σ-generalized rotations and auto-rotational invariance. The first result that we present provides intuition about the action of a Σ-generalized rotation, and how it relates to classical rotations.

Lemma 14 *Let* $\Sigma \succ 0$ *and* $B \in \mathbb{R}^{d\times d}$ *be such that*

$$B\Sigma + \Sigma B^\top = 0.$$

Then, $R = \Sigma^{-1/2} B \, \Sigma^{1/2}$ *is skew-symmetric, and for all* $\theta \in \mathbb{R}$,

$$e^{B\theta} = \Sigma^{1/2} \, e^{R\theta} \, \Sigma^{-1/2}.$$

In other words, the Σ*-generalized rotation* $e^{B\theta}$ *acts as a classical rotation* $e^{R\theta}$ *after the change of variables* $y = \Sigma^{-1/2}x$, *which rescales* $\mathbb{R}^d$ *along the principal axes of* Σ.

Proof Since $\Sigma \succ 0$, there exists a unique and invertible symmetric positive definite square root $\Sigma^{1/2}$. We then define

$$R = \Sigma^{-1/2} B \, \Sigma^{1/2}, \qquad R^\top = \Sigma^{1/2} B^\top \, \Sigma^{-1/2}.$$

Then, since $B\Sigma + \Sigma B^\top = 0$, we can multiply on the left by $\Sigma^{-1/2}$ and on the right by $\Sigma^{-1/2}$ to yield

$$0 = \Sigma^{-1/2} B \, \Sigma^{1/2} + \Sigma^{1/2} \, B^\top \, \Sigma^{-1/2} = R + R^\top,$$

which proves that R is skew-symmetric, and therefore $e^{R\theta}$ defines a classical rotation in $\mathbb{R}^d$, such that

$$e^{R\theta} = e^{\Sigma^{-1/2} B \Sigma^{1/2} \theta} = \Sigma^{-1/2} e^{B\theta} \Sigma^{1/2},$$

where the last equality can be seen by writing $e^{R\theta} = \sum_{n=0}^{\infty} \theta^n \frac{R^n}{n!}$ and noting that $R^n = (\Sigma^{-1/2} B \Sigma^{1/2})^n = \Sigma^{-1/2} B^n \Sigma^{1/2}$. Rearranging yields $e^{B\theta} = \Sigma^{1/2}\, e^{R\theta}\, \Sigma^{-1/2}$ as desired.

To see that the change of variables $y = \Sigma^{-1/2} x$ amounts to an anisotropic scaling, write the eigendecomposition $\Sigma = QDQ^\top$ with Q orthogonal and D diagonal with positive entries. Then $\Sigma^{-1/2} = QD^{-1/2}Q^\top$, i.e. the map rescales $\mathbb{R}^d$ by $D^{-1/2}$ along the principal axes of Σ, sending the covariance ellipsoid $\{x : x^\top \Sigma^{-1} x = 1\}$ to the unit sphere. ■

We now show that under auto-rotational invariance (Definition 2), the corresponding Σ that describes the geometry of the Σ-generalized rotations naturally corresponds to the distribution's second moment.

Lemma 15 *Let X be a r.v. with second moment $\Sigma = \mathbb{E}[XX^\top]$ and suppose that $e^{B\theta}X \sim X$ for all $\theta \in \mathbb{R}$. Then, $B\Sigma + \Sigma B^\top = 0$.*

Proof Let $C(\theta) = \mathbb{E}[(e^{B\theta}X)(e^{B\theta}X)^\top]$. $e^{B\theta}X \sim X$ implies that for all $\theta \in R$, $C(\theta)$ stays constant, such that $C(\theta) = \mathbb{E}[XX^\top] = \Sigma$. Directly computing $C(\theta)$ and its derivative yields

$$C(\theta) = e^{B\theta}\Sigma e^{B^\top \theta} = \Sigma$$
$$C'(\theta) = Be^{B\theta}\Sigma e^{B^\top\theta} + e^{B\theta}\Sigma B^\top e^{B^\top\theta} = 0$$

and we plug in $\theta = 0$ to obtain

$$C'(0) = B\Sigma + \Sigma B^\top = 0.$$

■

We note that the above derivation also follows through when we let $\Sigma = \mathrm{Cov}(X)$ and redefine $C(\theta)$ accordingly. It is however often more convenient to work with the non-centered geometry $\Sigma = \mathbb{E}[XX^\top]$. For example, we can show that if $\Sigma = \mathbb{E}[XX^\top] \succ 0$, the invariance confines the orbits of $e^{B\theta}$ to the ellipsoids determined by Σ, and forces the density to be constant on these orbits.

Lemma 16 *Let X be a random variable with $\Sigma = \mathbb{E}[XX^\top] \succ 0$ and continuous density p, and suppose $e^{B\theta}X \sim X$ for all $\theta \in \mathbb{R}$. Then p is constant along each orbit $\mathcal{O}_x = \{e^{B\theta}x : \theta \in \mathbb{R}\}$, and $\mathcal{O}_x$ lies inside the ellipsoid $\{y : y^\top \Sigma^{-1} y = x^\top \Sigma^{-1} x\}$.*

Proof First, we show that the orbit $\mathcal{O}_x$ lies inside the ellipsoid. As established in the proof of Lemma 15, we have $e^{B\theta}\Sigma e^{B^\top\theta} = \Sigma$ for every $\theta \in \mathbb{R}$. Inverting and replacing θ by $-\theta$ yields $e^{B^\top\theta}\Sigma^{-1}e^{B\theta} = \Sigma^{-1}$, hence

$$(e^{B\theta}x)^\top \Sigma^{-1}(e^{B\theta}x) = x^\top \Sigma^{-1} x \qquad \forall x \in \mathbb{R}^d,\ \theta \in \mathbb{R},$$

so $\mathcal{O}_x$ is contained in the ellipsoid $\{y : y^\top \Sigma^{-1} y = x^\top \Sigma^{-1} x\}$.

Then, to show that p is constant on $\mathcal{O}_x$, we take determinants in $e^{B\theta}\Sigma e^{B^\top\theta} = \Sigma$, which gives $(\det e^{B\theta})^2 = 1$. Hence $e^{B\theta}$ preserves Lebesgue measure, so $e^{B\theta}X$ has density $p \circ e^{-B\theta}$. Since $e^{B\theta}X \sim X$, and both densities are continuous, they agree everywhere, i.e.

$p(e^{B\theta}x) = p(x)$ for every $x \in \mathbb{R}^d$ and $\theta \in \mathbb{R}$. ■

We now classify select random variables that satisfy auto-rotational invariance. This includes all mean zero Gaussians and all random variables with degenerate second moment.

Lemma 17 *Let $X \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \succeq 0$. Then X is auto-rotationally invariant.*

Proof It suffices to consider $d \geq 2$, since all mean-zero Gaussians are auto-rotationally invariant in $d = 1$.

If Σ is rank-deficient, the claim follows from Lemma 18. Otherwise $\Sigma \succ 0$, and since $d \geq 2$ we may pick a nonzero skew-symmetric $C \in \mathbb{R}^{d\times d}$ and set $B = \Sigma^{1/2}C\Sigma^{-1/2} \neq 0$. Then

$$B\Sigma + \Sigma B^\top = \Sigma^{1/2}(C + C^\top)\Sigma^{1/2} = 0,$$

so $\frac{d}{d\theta}\big(e^{B\theta}\Sigma e^{B^\top\theta}\big) = 0$ and hence $e^{B\theta}\Sigma e^{B^\top\theta} = \Sigma$ for all θ. Since $e^{B\theta}X$ is Gaussian with mean 0 and covariance $e^{B\theta}\Sigma e^{B^\top\theta} = \Sigma$, we conclude $e^{B\theta}X \sim X$ for all θ. ■

Lemma 18 *Let X be a d-dimensional r.v. with rank-degenerate second moment $\Sigma = \mathbb{E}[XX^\top]$. Then*

$$e^{B\theta}X \sim X \quad \forall\theta \in \mathbb{R}$$

admits a non-trivial solution $B \neq 0$.

Proof Since Σ is rank-degenerate, it has a non-trivial kernel, and thus some $u \neq 0$ with $\Sigma u = 0$. Then, it follows that $\mathbb{E}[(u^\top X)^2] = u^\top\Sigma u = 0$, from which we conclude that $u^\top X = 0$ a.s.

To design our non-trivial solution B, we then pick $B = wu^\top$ for any $w \neq 0$. Then $BX = w(u^\top X) = 0$ a.s., so $e^{B\theta}X = X$ a.s. for every θ, and in particular $e^{B\theta}X \sim X$. ■

The following result demonstrates the connection between weak auto-rotational invariance (Definition 4) and full auto-rotational invariance (Definition 2).

Lemma 19 *Let X be a d-dimensional random variable with bounded support. Then X is weakly auto-rotationally invariant if and only if it is auto-rotationally invariant.*

Proof Full ARI (Definition 2) implies weak ARI (Definition 4) since equality of the law of $e^{B\theta}X$ to the law of X, implies that all their cumulants agree.

To prove the converse, suppose X is weakly ARI with matrix $B \neq 0$. Since the support is bounded, the moment generating function $M(v) = \mathbb{E}[e^{v^\top X}]$ is entire and positive, so the cumulant generating function $K = \log M$ is real-analytic on $\mathbb{R}^d$ with $K(v) = \sum_{n\geq 1}\frac{1}{n!}T_n^X(v,\ldots,v)$ near the origin. By the definition of weak ARI, every cumulant of order $n \neq 2$ is invariant under the map $v \mapsto e^{B^\top\theta}v$. Since the cumulant generating function of $e^{B\theta}X$ evaluated at v is $K(e^{B^\top\theta}v)$, it follows that the cumulant generating functions of X and $e^{B\theta}X$ differ only in their quadratic terms, i.e., their covariances:

$$K(e^{B^\top\theta}v) - K(v) = \tfrac{1}{2}v^\top Q_\theta v, \qquad Q_\theta = e^{B\theta}\mathrm{Cov}(X)e^{B^\top\theta} - \mathrm{Cov}(X). \tag{46}$$

This equality holds not just near the origin, but actually on all of $\mathbb{R}^d$ since both sides are real-analytic. We now show that the quadratic difference Q_θ is in fact 0.

Let $h(v) = \sup\{v^\top x : x \in \text{supp}(X)\}$, which is finite since the support is bounded. Then $K(sv) \leq s\, h(v)$ for $s > 0$, while $K(sv) \geq s\, v^\top \mathbb{E}[X]$ by Jensen's inequality, so $K(sv) = O(s)$ along every ray. Applying this to both terms on the left of (46) gives $\frac{s^2}{2} v^\top Q_\theta v = O(s)$. Thus, if we divide by s^2 and let $s \to \infty$, then $v^\top Q_\theta v = 0$ for all v, i.e. $Q_\theta = 0$. From this, we conclude that $K(e^{B^\top \theta} v) = K(v)$ for all v and θ, so we have full auto-rotational invariance: $e^{B\theta} X \sim X$. ■

A simple class of random variables that fail weak auto-rotational invariance (which we use in our numerical experiments) are discrete random variables, whose support spans the full space $\mathbb{R}^d$.

Proposition 20 *Let $X_0 \sim p_0$ be a discrete random variable with bounded support. Then X_0 is weakly auto-rotationally invariant if and only if the support of p_0 fail to span $\mathbb{R}^d$.*

Proof If support fails to span $\mathbb{R}^d$, then some $u \neq 0$ satisfies $u^\top X_0 = 0$ a.s., so $\Sigma = \mathbb{E}[X_0 X_0^\top]$ is rank-degenerate and by Lemma 18, X_0 is auto-rotationally invariant, and hence weakly auto-rotationally invariant.

Conversely, suppose for contradiction that the support spans $\mathbb{R}^d$ and that X_0 is weakly ARI with matrix $B \neq 0$. We note that since the support is bounded, then, by Lemma 19, it suffices to derive a contradiction from the fact that X_0 satisfies full ARI while also having a support that spans $\mathbb{R}^d$. Since $e^{B\theta} X_0 \sim X_0$ for all θ, the map $e^{B\theta}$ must permute the atoms of p_0. For any atom x, the orbit $\{e^{B\theta} x : \theta \in \mathbb{R}\}$ must therefore consist of atoms. But the orbit is a connected set, since it is a continuous image of $\mathbb{R}$. Since the set of atoms is countable, and a connected subset of $\mathbb{R}^d$ with more than one point is uncountable, it follows that the orbit is a single point, i.e. $e^{B\theta} x = x$ for all θ. Differentiating at $\theta = 0$ then gives $Bx = 0$. Since this holds for every atom and the atoms span $\mathbb{R}^d$, we conclude $B = 0$, which contradicts the ARI assumption that $B \neq 0$. ■

Appendix C. Additional Proofs

Lemma 21 *If the covariances of two linear additive noise SDEs match for all times, i.e. $\text{Cov}(X_t) = \text{Cov}(Y_t)\ \forall\ t \geq 0$, such that X_t evolves with parameters (A, H) and Y_t evolves with parameters $(\tilde{A}, \tilde{H})$, then their residual parameters $\bar{A} = A - \tilde{A}$ and $\bar{H} = H - \tilde{H}$ obey the Lyapunov equation*

$$\bar{A} C_t + C_t \bar{A}^\top + \bar{H} = 0. \tag{47}$$

Thus, for any distinct times $t_1 \neq t_2$, we have

$$\bar{A}(C_{t_2} - C_{t_1}) + (C_{t_2} - C_{t_1})\bar{A}^\top = 0. \tag{48}$$

Proof Recall that given SDE parameters (A, H), the covariance $C_t = \mathrm{Cov}(X_t)$ obeys

$$C_t = e^{At} C_0 e^{A^\top t} + \int_0^t e^{A(t-s)} H e^{A^\top (t-s)} ds. \tag{49}$$

Differentiating (49) shows that the covariance of each system obeys its Lyapunov ODE, so for every $t \geq 0$:

$$\frac{d}{dt} C_t = A C_t + C_t A^\top + H \tag{50}$$

$$\frac{d}{dt} C_t = \tilde{A} C_t + C_t \tilde{A}^\top + \tilde{H}. \tag{51}$$

Since the covariances of X_t and Y_t are equal at all time points, so are their derivatives. Subtracting yields the desired equation (47) at every $t \geq 0$. ■

Proposition 22 *For any initial covariance $C_0 \in \mathrm{Sym}_d$, any $m \geq 1$, and any distinct times $t_1, \ldots, t_D > 0$ (where $D = \frac{d(d+1)}{2}$), there exist parameters (A^*, G^*) such that $\det(M(A^*, G^*)) \neq 0$.*

Proof Let $A^* = \mathrm{diag}(-2^1, -2^2, \ldots, -2^d)$. Since A^* is diagonal, the covariance ODE decouples entry-by-entry:

$$\dot{C}_{ij} = \mu_{ij}\, C_{ij} + (GG^\top)_{ij}, \qquad \mu_{ij} = \lambda_i + \lambda_j = -2^i - 2^j.$$

The rates $\{\mu_{ij}\}_{i \leq j}$ are all distinct, since distinct subsets of $\{2^1, \ldots, 2^d\}$ have distinct sums. The scalar solution gives

$$\Delta C_k^{(ij)} = C_{ij}(t_k) - C_{ij}(0) = \alpha_{ij}\big(e^{\mu_{ij}\, t_k} - 1\big),$$

where $\alpha_{ij} = C_{ij}(0) + (GG^\top)_{ij}/\mu_{ij}$. Ordering the D pairs (i, j) with $i \leq j$ as $\ell = 1, \ldots, D$, the matrix M factors as

$$M = F \cdot \mathrm{diag}(\alpha_\ell)_{\ell=1}^D,$$

where $F_{k\ell} = e^{\mu_\ell\, t_k} - 1$. Therefore,

$$\det(M) = \det(F) \cdot \prod_{\ell=1}^{D} \alpha_\ell. \tag{52}$$

We show that both factors are nonzero. For the first factor, consider the $D+1$ functions $\{1, e^{\mu_1 t}, \ldots, e^{\mu_D t}\}$ evaluated at the $D+1$ distinct points $\{0, t_1, \ldots, t_D\}$. Since the exponents $0, \mu_1, \ldots, \mu_D$ are all distinct, these functions form a Chebyshev system (Karlin and Studden, 1966, Ch. 1), meaning that their evaluation matrix at any set of distinct points is non-singular. In particular, the $(D{+}1) \times (D{+}1)$ matrix

$$\widetilde{V} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{\mu_1 t_1} & \cdots & e^{\mu_D t_1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{\mu_1 t_D} & \cdots & e^{\mu_D t_D} \end{bmatrix}$$

is non-singular. Subtracting the first row from each subsequent row gives

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & e^{\mu_1 t_1} - 1 & \cdots & e^{\mu_D t_1} - 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & e^{\mu_1 t_D} - 1 & \cdots & e^{\mu_D t_D} - 1 \end{bmatrix},$$

whose determinant equals $\det(F)$ by cofactor expansion along the first column. Since row operations preserve non-singularity, $\det(F) \neq 0$ for any choice of distinct $t_k > 0$.

Then, to ensure that the second factor is nonzero, we take $G^* = \sqrt{c}\,\mathbf{1}e_1^\top$, where $\mathbf{1} \in \mathbb{R}^d$ is the vector of all 1's, $e_1 \in \mathbb{R}^m$ is the first standard basis vector, and $c > 0$. Since every entry of $G^*(G^*)^\top$ equals c, we have $\alpha_{ij} = C_{ij}(0) - c/(2^i + 2^j)$, which vanishes for at most one value of c for each of the D pairs (i, j). Choosing $c > 0$ outside these finitely many values gives $\alpha_\ell \neq 0$ for all ℓ. Thus, by (52), $\det(M(A^*, G^*)) \neq 0$, as desired.

■

Appendix D. SDEs and Causality

Although nonzero entries in the observational diffusion $H = GG^\top$ inform the presence of latent confounders between pairs of variables, we note that it is impossible for H alone to determine latent confounders over $p > 2$ endogenous variables, $X_t^{(i_1)}, \cdots, X_t^{(i_p)}$. Indeed, given only $H = GG^\top$, multiple causal interpretations may be possible, because structurally different matrices $G, \tilde{G}$ can obey $H = GG^\top = \tilde{G}\tilde{G}^\top$.

Example 4 (Non-identifiability of higher order latent confounders) *Consider two matrices G_1, G_2, which share the same observational diffusion H:*

$$G_1 = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \quad G_2 = \begin{bmatrix} \frac{4}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{4}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{4}{3} \end{bmatrix}, \quad H = G_1 G_1^\top = G_2 G_2^\top = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

Given G_1 as the additive noise of an SDE, each pair of variables shares a noise source, since each row shares a nonzero column entry with another row. However, there is no driving noise source that is shared among all three variables, since each column contains a 0. In contrast, the causal interpretation under G_2 represents driving noise shared by all three variables.

Since G_1 and G_2 admit different causal graphs, despite having the same observational diffusion H, this shows that H provides information about the existence of unobserved confounders between pairs of variables, but cannot provide further causal structure about the confounder with respect to the other endogenous variables $[d] \setminus \{i, j\}$.

Similarly, we can have $H_{i,j} = 0$, while variables $X^{(i)}$ and $X^{(j)}$ share a noise source $W_t^{(k)}$.

Example 5 (Zeros in diffusion H can result from pairwise latent confounders)

$$G = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \implies H = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

From G, $X^{(1)}$ and $X^{(2)}$ both depend on $W^{(1)}$ and $W^{(2)}$, however, this cannot be determined from the observational diffusion H, due to cancellations.

We note that if the noise is not additive, then this further complicates the causal interpretation. In this case, the driving noise $\sigma(X_t)$ may be a function of the endogenous variables, i.e. $\beta(j) \cap V \neq \emptyset$. Thus, unlike the additive noise setting in Lemma 9, directed edges $i \to j$ may be informed by the driving noise, via $\beta(j)$, rather than just the drift, via $\alpha(j)$. However, since only $\sigma(X_t)\sigma(X_t)^\top$ is observable rather than $\sigma(X_t)$ itself, observational data under such a model would lend itself to multiple interpretations of the causal graph. The idea is similar to Example 4, where we saw one interpretation feature three pairwise latent confounders and another feature a single confounder over all variables, but under non-additive noise, the ambiguity extends to edges $i \to j$ within the endogenous set. This is illustrated in Example 5.5 in Hansen and Sokol (2014).

The causal framework also highlights the impracticality of common model assumptions within the scRNA-seq literature, namely isotropic diffusion $H = \sigma^2 \mathrm{I}_d$ and gradient-field drift $\nabla\psi$. Except in exceedingly rare cases of complete cancellations when taking $H = GG^\top$ (see Example 5), isotropic diffusion cannot model latent confounders from shared noise. Moreover, while we expect GRNs to contain feedback loops, i.e. cycles in $\mathcal{G}$, imposing irrotational drift prevents most cycles from being considered. Lemma 23 shows that imposing irrotationality on a linear model is equivalent to imposing a symmetric drift, which determines a symmetric causal graph. However, as illustrated in Weinreb et al. (2018, Fig. 6), symmetric GRNs cannot capture important relationships, such as negative feedback loops or repressilator dynamics. Indeed, consider the 2-cycle $i \leftrightarrow j$. By Lemma 9, $i \to j$ if $A_{j,i} \neq 0$ and $j \to i$ if $A_{i,j} \neq 0$. However, if the model is irrotational, then by Lemma 23, $A_{i,j} = A_{j,i}$, which can only model a positive feedback loop.

Lemma 23 *Let X_t evolve according to a linear additive noise SDE* (3)*. Then, the drift AX_t is irrotational if and only if A is symmetric, i.e. $A = A^\top$.*

Proof First, we note that for any symmetric matrix A, the vector field Ax can be expressed as the gradient of the scalar potential given by the quadratic form $\phi(x) = x^\top Ax/2$ (Petersen et al., 2008, (96)-(97)). Hence, Ax is irrotational if $A = A^\top$.

For the opposite direction, we recall the fact that any gradient field from $\mathbb{R}^n \to \mathbb{R}^n$ has a symmetric Jacobian (Williamson and Trotter, 1974; Campbell, 2014). Since the Jacobian of a linear vector field $Ax : \mathbb{R}^n \to \mathbb{R}^n$ is A, we conclude that all irrotational linear vector fields from $\mathbb{R}^n \to \mathbb{R}^n$ are given by symmetric matrices. ■

Appendix E. Entropy-Regularized Optimal Transport

The entropy-regularized optimal transport problem also admits a dual formulation, in terms of finding a pair of potentials (f, g) with respect to the Gaussian transition kernel $K(x, y) \propto$

$e^{\frac{-c(x,y)}{\epsilon^2}} \propto e^{\frac{-\|y-x\|^2}{2\epsilon^2}}$ (Janati, 2021; Nutz, 2021):

$$\pi^*(x,y) = e^{f^*(x)+g^*(y)} K(x,y)\,\mathrm{d}\mu(x)\mathrm{d}\nu(y) \tag{53}$$

$$f^*, g^* = \underset{f\in\mathcal{L}^1(\mu), g\in\mathcal{L}^1(\nu)}{\arg\max} \mathbb{E}_\mu(f) + \mathbb{E}_\nu(g) - \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{f(x)+g(y)} K(x,y)\,\mathrm{d}\mu(x)\mathrm{d}\nu(y) - 1\right). \tag{54}$$

In particular, given marginals μ, ν and the transition kernel K, Sinkhorn's algorithm uses the dual formulation (54) to find the ϵ-entropy regularized optimal transport solution π^*, by solving for f^*, g^* via iterative projections (Peyré et al., 2019).

Definition 24 (Markov concatenation of couplings) *Given Polish spaces X_1, X_2, X_3 and couplings $\pi_{1,2} \in \mathcal{P}(X_1 \times X_2)$ and $\pi_{2,3} \in \mathcal{P}(X_2 \times X_3)$ with identical marginals μ_2 on X_2, the* Markov concatenation $\pi_{2,3} \circ \pi_{1,2}$ *is a multi-coupling in $\mathcal{P}(X_1 \times X_2 \times X_3)$ given by*

$$(\pi_{2,3} \circ \pi_{1,2})(\mathrm{d}x_1, \mathrm{d}x_2, \mathrm{d}x_3) = \pi_{1,2}(\mathrm{d}x_1 \mid x_2)\mu_2(\mathrm{d}x_2)\pi_{2,3}(\mathrm{d}x_3 \mid x_2).$$

Here $\pi_{1,2}(\mathrm{d}x_1 \mid x_2)$ is the disintegration of $\pi_{1,2}$ with respect to μ_2 and $\pi_{2,3}(\mathrm{d}x_3 \mid x_2)$ is the disintegration of $\pi_{2,3}$ with respect to μ_2.

The interpretation of the Markov concatenation is as follows: a random "trajectory" according to $\pi_{23} \circ \pi_{12}$ corresponds to taking the first two steps distributed according to π_{12}, then the third step distributed according to "π_{23} conditional on the second marginal of π_{12}". The existence of the Markov concatenation is guaranteed by the disintegration theorem, and Markov concatenations appear naturally in the time-discretized version of trajectory inference via Schrödinger bridges (Lavenant et al., 2021). In particular, given Polish spaces $X_1, \ldots, X_4$ and couplings $\pi_{12}, \pi_{23}, \pi_{34}$, it holds that Markov concatenation is associative, and so we can unambiguously define the iterated Markov concatenation $\pi_{34} \circ \pi_{23} \circ \pi_{12}$ (see Benamou et al. (2019) Section 3.2).

Proposition 25 (Optimality of anisotropic entropic optimal transport) *Let the reference SDE be a linear additive noise SDE with drift A and diffusion H (3). Given a set of marginals $p_0, \ldots, p_{N-1}$ over times $\{t_i\}_{i=0}^{N-1}$ (not necessarily coming from the reference SDE) for which $D_{KL}(p_i \mid Leb) < \infty$ for all $1 \le i \le N-1$, and $D_{KL}(p_i \mid (p_{A,H})_i) < \infty$ for all $1 \le i \le N-2$, suppose there exists $\pi_{i,i+1}$, the anisotropic entropic OT solution (31) obtained by Sinkhorn's algorithm with marginals $\mu = p_i, \nu = p_{i+1}$ and transition kernel*

$$K^i_{A,H}(x,y) \propto \exp(-\frac{1}{2}(y - e^{A(t_{i+1}-t_i)}x)^\top(\Sigma_i)^{-1}(y - e^{A(t_{i+1}-t_i)}x)) \tag{55}$$

where $\Sigma_i = \int_{t_i}^{t_{i+1}} e^{A(t_{i+1}-s)} H e^{A^\top(t_{i+1}-s)}\,\mathrm{d}s$. Let π denote the joint distribution given by the Markov concatenation of couplings (see Definition 24):

$$\pi = \pi_{N-2,N-1} \circ \ldots \circ \pi_{0,1}.$$

Then, π minimizes relative entropy to the law of the reference SDE among Markov couplings:

$$\pi = \underset{\pi\in\Pi(p_0,\ldots,p_{N-1}),\ \pi\ Markov}{\arg\min} D_{KL}(\pi \| p^N_{A,H}),$$

where $p^N_{A,H}$ is the law of the reference SDE, discretized over time points $t_0, \ldots, t_{N-1}$.

Proof For each $i = 0, \ldots, N-2$, by (5), we have $\pi_{i,i+1} = \arg\min_{\pi \in \Pi(p_i, p_{i+1})} D_{KL}(\pi \| K^i_{A,H})$. Since $K^i_{A,H} \sim \mathcal{N}(e^{A(t_{i+1}-t_i)}X_t, \Sigma_i)$ is the transition kernel of the SDE at time t_i (Särkkä and Solin, 2019, Sec 6.1), this implies that

$$\pi_{i,i+1} = \underset{\pi \in \Pi(p_i, p_{i+1})}{\arg\min} D_{KL}(\pi \| (p_{A,H})_{i,i+1}),$$

where $(p_{A,H})_{i,i+1}$ denotes the joint distribution of the marginals of $p_{A,H}$ at the two time points t_i, t_{i+1}. The details are as follows. Let $K_{i,i+1}(x, y)$ denote the transition kernel from time t_i to time t_{i+1} for the SDE $\mathrm{d}X_t = AX_t\,\mathrm{d}t + G\mathrm{d}W_t$. In particular, if $(p_{A,H})_i$ is the marginal of this SDE at time t_i, then we have that $K_{i,i+1}(x, y)\mathrm{d}(p_{A,H})_i(x)\mathrm{d}y$ is equal to the joint distribution $(p_{A,H})_{i,i+1}$. Hence

$$D_{KL}(\pi \mid (p_{A,H})_{i,i+1}) = D_{KL}(\pi \mid K_{A,H}(x, y)d(p_{A,H})_i(x)dy).$$

We claim that for any $\pi \in \Pi(p_i, p_{i+1})$,

$$\begin{aligned}
D_{KL}(\pi \mid K_{A,H}(x, y)d(p_{A,H})_i dy) &= \int \log\left(\frac{d\pi}{d(K \cdot (p_{A,H})_i \otimes \mathrm{Leb})}(x, y)\right) d\pi \\
&= \int \log\left(\frac{d\pi}{d(K \cdot p_i \otimes dp_{i+1})}(x, y)\frac{d(K \cdot p_i \otimes dp_{i+1})}{d(K \cdot (p_{A,H})_i \otimes \mathrm{Leb})}(x, y)\right) d\pi \\
&= \int \log\left(\frac{d\pi}{d(K \cdot p_i \otimes dp_{i+1})}(x, y)\frac{dp_i}{d(p_{A,H})_i}(x)\frac{dp_{i+1}}{d\mathrm{Leb}}(y)\right) d\pi \\
&= D_{KL}(\pi \mid K_{A,H}(x, y)dp_i(x)dp_{i+1}(y)) + D_{KL}(p_i \mid (p_{A,H})_i) \\
&\quad + D_{KL}(p_{i+1} \mid \mathrm{Leb}).
\end{aligned}$$

Note that $(p_{A,H})_0 = p_0$. Hence, under the assumption that for all $i = 1, \ldots, N-1$, $D_{KL}(p_i \mid \mathrm{Leb}) < \infty$, the minimizers for the following two minimization problems are identical:

$$\min_{\pi \in \Pi(p_i p_{i+1})} D_{KL}(\pi \mid (p_{A,H})_{i,i+1}) \text{ and } \min_{\pi \in \Pi(p_i p_{i+1})} D_{KL}(\pi \mid K_{A,H}(x, y)dp_i(x)dp_{i+1}(y)).$$

Since π is constructed as a Markov concatenation, the conclusion follows from Lemma 3.4 of Benamou et al. (2019), which in this case tells us that: if $p^N_{A,H}$ is the projection of the law of the SDE onto the set of times $\{t_0, \ldots, t_{N-1}\}$, then for any N-coupling π which is constructed as a Markovian concatenation $\pi_{N-2,N-1} \circ \ldots \pi_{0,1}$, and has ith marginal equal to p_i, we have

$$D_{KL}(\pi \mid p^N_{A,H}) = \sum_{i=0}^{N-2} D_{KL}(\pi_i \mid (p_{A,H})_{i,i+1}) - \sum_{i=1}^{N-2} D_{KL}(p_i \mid (p_{A,H})_i).$$

Hence minimizing over each $D_{KL}(\pi_{i,i+1} \mid (p_{A,H})_{i,i+1})$ (for $\pi \in \Pi(p_i, p_{i+1})$) is equivalent to minimizing over Markovian π's with ith marginal p_i.

■

The previous proposition assumes that $D_{KL}(p_i \mid \mathrm{Leb}) < \infty$ for all $i \neq 0$. The following computation shows that this assumption is satisfied if the diffusion matrix is non-degenerate.

We note that $X_t \sim e^{At}X_0 + \int_0^t e^{A(t-s)}GdW_s$. As X_0 is independent from the noise W_s, and $\int_0^t e^{A(t-s)}GdW_s \sim \mathcal{N}(0, \Sigma_t)$, where $\Sigma_t = \int_0^t e^{A(t-s)}He^{A^\top(t-s)}ds$, it follows that

$$p_t = e^{At}p_0 * \mathcal{N}(0, \Sigma_t).$$

Thus, we may compute (Cover and Thomas, 1991, Theorem 9.4.1)

$$D_{KL}(\mathcal{N}(0, \Sigma_t) \mid Leb) = -\frac{1}{2}\log((2\pi e)^d \det(\Sigma_t)).$$

From the definition of Σ_t, $\det(\Sigma_t) \neq 0$ as long as $H = GG^\top$ is full rank. Then, we use the fact that the $D_{KL}(\cdot \mid Leb)$ is convex, and apply Jensen's inequality for measures, to deduce that

$$D_{KL}(p_0 * \mathcal{N}(0, \Sigma_t) \mid Leb) \leq \int D_{KL}(\mathcal{N}(x, \Sigma_t) \mid Leb)dp_0(x) = D_{KL}(\mathcal{N}(0, \Sigma_t) \mid Leb).$$

This shows directly that p_t has finite entropy whenever $t > 0$ as desired.

Remark 26 (Applying EOT for SDE trajectory inference) *In the standard entropic OT problem* (30)*, the cost is the squared Euclidean distance, and $K \sim \mathcal{N}(x, \epsilon^2 \mathrm{I}_d)$ is an isotropic Gaussian kernel. As noted in Lavenant et al. (2021), this implies that entropy regularized OT can be leveraged for trajectory inference from observed marginals, given a reference SDE. For example, to find the discretized law on paths $\pi^* \in \Pi(\mu, \nu)$ satisfying $P(X_t = \mu, X_{t+\Delta t} = \nu)$, which minimizes relative entropy to the law of a pure diffusion process $\mathrm{d}X_t = \sigma \mathrm{d}W_t$, one should set the entropic regularization $\epsilon^2 = \sigma^2 \Delta t$. This would correspond to minimizing the KL divergence to $K(x, y) = e^{-\frac{\|y-x\|^2}{2\sigma^2\Delta t}} \sim \mathcal{N}(x, \sigma^2\Delta t)$, which is the approximate transition kernel of the reference SDE.*

Appendix F. Maximum Likelihood Estimation

Proof [Proof of Proposition 11] We follow the standard procedure for deriving maximum likelihood estimators (Peiris et al., 2003; Pavliotis, 2014). Our likelihood function is given by

$$\mathcal{L} = \Pi_{i=0}^{N-2}\left(\Pi_{j=0}^{M-1} p(X_{i+1}^{(j)} | X_i^{(j)})\right),$$

where we have denoted $X_i = X_{t_i}$ for shorthand. As in Pavliotis (2014), we consider the discretized law $X_{i+1}|X_i \sim \mathcal{N}(X_i + AX_i\Delta t, H\Delta t)$, which implies

$$p(X_{i+1}^{(j)} | X_i^{(j)}) = \frac{1}{(2\pi\Delta t)^{d/2}\det(H)^{1/2}} \exp\left(-\frac{1}{2}(\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)^\top (H\Delta t)^{-1}(\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)\right),$$

where $\Delta X_i^{(j)} = X_{i+1}^{(j)} - X_i^{(j)}$. Plugging this back into the likelihood expression yields

$$\begin{aligned}
\mathcal{L} &= \frac{1}{(2\pi\Delta t)^{\frac{dM(N-1)}{2}} \det(H)^{\frac{M(N-1)}{2}}} \\
&\quad \times \exp\Big(-\sum_{i=0}^{N-2}\sum_{j=0}^{M-1} \frac{1}{2}(\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)^\top (H\Delta t)^{-1}(\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)\Big) \\
\log(\mathcal{L}) &= -\frac{M(N-1)}{2}\big(d\log(2\pi\Delta t) + \log(\det(H))\big) \\
&\quad - \frac{1}{2}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1} \Big((\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)^\top (H\Delta t)^{-1}(\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)\Big).
\end{aligned}$$

We then derive the maximum likelihood estimators through matrix calculus (Petersen et al., 2008):

$$\begin{aligned}
\frac{d\log(\mathcal{L})}{dA} &= -\frac{1}{2}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1} \frac{d}{dA}\Big((\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)^\top (H\Delta t)^{-1}(\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)\Big) \\
&= -\frac{1}{2}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1} -2\Delta t\frac{d}{dA}\Big((\Delta X_i^{(j)})^\top (H\Delta t)^{-1}AX_i^{(j)}\Big) + \Delta t^2\frac{d}{dA}\Big({X_i^{(j)}}^\top A^\top (H\Delta t)^{-1}AX_i^{(j)}\Big) \\
&= -\frac{1}{2}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1} -2\Delta t(H\Delta t)^{-1}(\Delta X_i^{(j)}){X_i^{(j)}}^\top + 2\Delta t^2(H\Delta t)^{-1}AX_i^{(j)}{X_i^{(j)}}^\top \\
&= \sum_{i=0}^{N-2}\sum_{j=0}^{M-1} (H)^{-1}(\Delta X_i^{(j)}){X_i^{(j)}}^\top - \Delta t(H)^{-1}AX_i^{(j)}{X_i^{(j)}}^\top.
\end{aligned}$$

We can solve for the MLE linear drift A by setting $\frac{d\log(\mathcal{L})}{dA} = 0$:

$$\begin{aligned}
\Delta t(H)^{-1}A\sum_{i=0}^{N-2}\sum_{j=0}^{M-1} X_i^{(j)}{X_i^{(j)}}^\top &= (H)^{-1}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1} \Delta X_i^{(j)}{X_i^{(j)}}^\top \\
A &= \frac{1}{\Delta t}\left(\sum_{i=0}^{N-2}\sum_{j=0}^{M-1} \Delta X_i^{(j)}{X_i^{(j)}}^\top\right)\left(\sum_{i=0}^{N-2}\sum_{j=0}^{M-1} X_i^{(j)}{X_i^{(j)}}^\top\right)^{-1}.
\end{aligned}$$

We now estimate the diffusion H. For simplicity, we work with $H^{-1} = (GG^\top)^{-1}$

$$\begin{aligned}
\frac{d\log(\mathcal{L})}{dH^{-1}} &= \frac{d}{dH^{-1}}\Big(\frac{M(N-1)}{2}\log(\det(H^{-1})) \\
&\qquad - \frac{1}{2\Delta t}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}\Big((\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)^\top H^{-1}(\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)\Big)\Big) \\
&= \frac{M(N-1)}{2}H - \frac{1}{2\Delta t}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}\Big((\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)(\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)^\top\Big).
\end{aligned}$$

We can solve for the MLE of H by setting $\frac{d\log(\mathcal{L})}{dH^{-1}} = 0$

$$H = \frac{1}{M(N-1)\Delta t}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}\left((\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)(\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)^\top\right)$$
$$= \frac{1}{MT}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}\left((\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)(\Delta X_i^{(j)} - AX_i^{(j)}\Delta t)^\top\right).$$

■

Corollary 27 *Let X_t evolve according to a general additive noise SDE (2). Then the Euler-approximated likelihood function is given by*

$$\mathcal{L} = \frac{1}{(2\pi\Delta t)^{\frac{dM(N-1)}{2}}\det(H)^{\frac{M(N-1)}{2}}}$$
$$\times \exp\left(-\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}\frac{1}{2}(\Delta X_i^{(j)} - b(X_i^{(j)})\Delta t)^\top(H\Delta t)^{-1}(\Delta X_i^{(j)} - b(X_i^{(j)})\Delta t)\right).$$

Taking the logarithm gives

$$\log(\mathcal{L}) = -\frac{M(N-1)}{2}\left(d\log(2\pi\Delta t) + \log(\det(H))\right)$$
$$-\frac{1}{2}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}\left((\Delta X_i^{(j)} - b(X_i^{(j)})\Delta t)^\top(H\Delta t)^{-1}(\Delta X_i^{(j)} - b(X_i^{(j)})\Delta t)\right).$$

Hence, given that the drift function b is parameterized by values $\alpha_b^{(k)}$, the maximum likelihood solution for the drift function b obeys

$$0 = -\frac{1}{2}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}\frac{d}{d\alpha_b^{(k)}}\left((\Delta X_i^{(j)} - b(X_i^{(j)})\Delta t)^\top(H\Delta t)^{-1}(\Delta X_i^{(j)} - b(X_i^{(j)})\Delta t)\right).$$

The MLE for the diffusion H admits the closed solution

$$H = \frac{1}{MT}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}\left((\Delta X_i^{(j)} - b(X_i^{(j)})\Delta t)(\Delta X_i^{(j)} - b(X_i^{(j)})\Delta t)^\top\right).$$

Remark 28 *Previous works have predominantly focused on the case of observing a single long trajectory rather than a collection of short trajectories. Suppose that one observes a set of M observed trajectories, then drift estimation may also be performed by averaging the classical MLE estimator for a single trajectory across observations. In dimension 1, this is given by $\hat{A}_T = \frac{1}{M}\sum_j \frac{\int_0^T X_t^{(j)}\,\mathrm{d}X_t^{(j)}}{\int_0^T (X_t^{(j)})^2\,\mathrm{d}t}$. Note that this is distinct from the MLE estimator that we derived for multiple trajectories in (34). Indeed, we observe that the latter estimator converges at a much faster rate than the averaged classical estimator.*

Appendix G. Additional Experiments

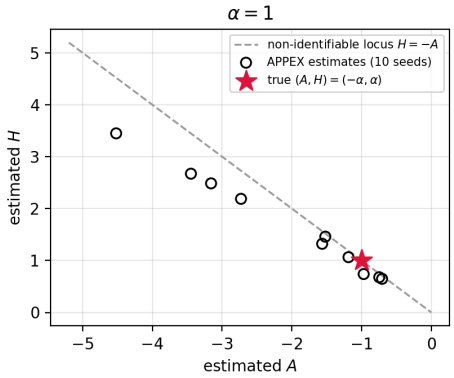
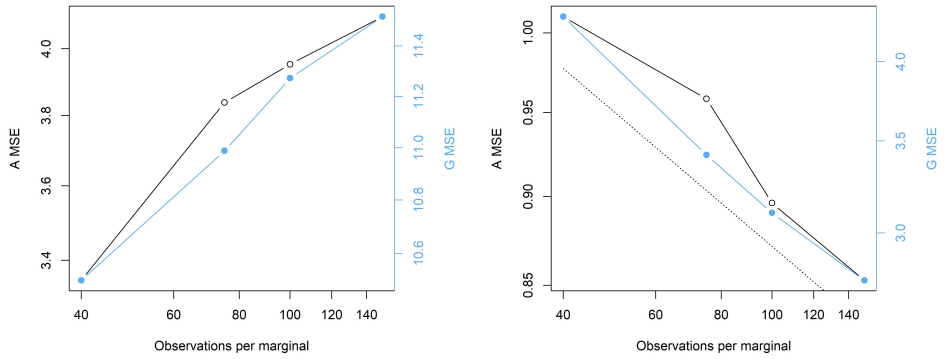


Figure 8: We plot APPEX's drift and diffusion estimates across 10 random seeds for the Ornstein–Uhlenbeck SDE $\mathrm{d}X_t = -X_t\,\mathrm{d}t + \mathrm{d}W_t$ from the non-identifiable setting, where marginals are sampled from $p_{\mathrm{eq}} = \mathcal{N}(0, \frac{1}{2})$.

We also present the results of additional experiments verifying the consistency of our method, as we increase the number of samples per marginal M (Figures 9 and 10).

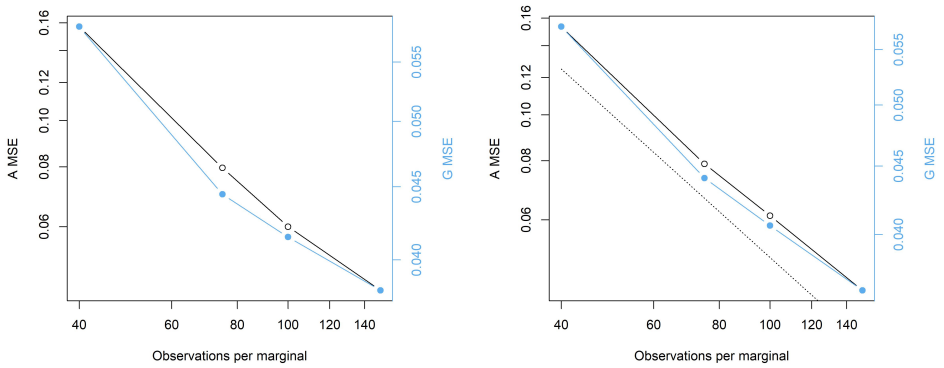


(a) WOT (b) APPEX

Figure 9: Average mean squared error (MSE) of A and $GG^\top$ as the number of observations per marginal increases across 50 random 2d systems. Note that this is a log-log plot and the dotted line in the APPEX plot shows that the convergence follows a power law relationship with exponent -0.125. The number of observed marginals is fixed at 20.

References

A. Aalto, L. Viitasaari, P. Ilmonen, L. Mombaerts, and J. Gonçalves. Gene regulatory network inference from sparsely sampled noisy data. *Nature communications*, 11(1):3493, 2020.



(a) WOT (b) APPEX

Figure 10: Average mean squared error (MSE) of A and $GG^\top$ as the number of observations per marginal increases across 70 random 12d systems. In this log-log plot, the dotted line in the APPEX plot indicates that the convergence follows a power law relationship with exponent -1. The number of observed marginals is fixed at 20.

E. E. Adams and L. W. Gelhar. Field study of dispersion in a heterogeneous aquifer: 2. spatial moments analysis. *Water Resources Research*, 28(12):3293–3307, 1992.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

Aristotle. *Generation of Animals*, volume 366 of *Loeb Classical Library*. Harvard University Press, Cambridge, MA, 1942.

L. Atanackovic, A. Tong, B. Wang, L. J. Lee, Y. Bengio, and J. S. Hartford. Dyngfn: Towards bayesian inference of gene regulatory networks with gflownets. *Advances in Neural Information Processing Systems*, 36, 2024.

J.-D. Benamou, G. Carlier, S. Di Marino, and L. Nenna. An entropy minimization approach to second-order variational mean-field games. *Mathematical Models and Methods in Applied Sciences*, 29(08):1553–1583, 2019.

K. J. Beven, D. E. Henderson, and A. D. Reeves. Dispersion parameters for undisturbed partially saturated soil. *Journal of hydrology*, 143(1-2):19–43, 1993.

J. P. Bishwal. *Parameter estimation in stochastic differential equations*. Springer, 2007.

F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of political economy*, 81(3):637–654, 1973.

P. Boeken and J. M. Mooij. Dynamic structural causal models. *arXiv preprint arXiv:2406.01161*, 2024.

V. I. Bogachev, N. V. Krylov, M. Röckner, and S. V. Shaposhnikov. *Fokker–Planck–Kolmogorov Equations*, volume 207. American Mathematical Society, 2022.

J. M. Boggs, S. C. Young, L. M. Beard, L. W. Gelhar, K. R. Rehfeldt, and E. E. Adams. Field study of dispersion in a heterogeneous aquifer: 1. overview and site description. *Water Resources Research*, 28(12):3281–3291, 1992.

L. Brogat-Motte, R. Bonalli, and A. Rudi. Learning controlled stochastic differential equations. *arXiv preprint arXiv:2411.01982*, 2024.

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

A. G. Cahill, R. Beckie, B. Ladd, E. Sandl, M. Goetz, J. Chao, J. Soares, C. Manning, C. Chopra, N. Finke, et al. Advancing knowledge of gas migration and fugitive gas from energy wells in northeast british columbia, canada. *Greenhouse Gases: Science and Technology*, 9(2):134–151, 2019.

L. Campbell. Reduction theorems for the strong real jacobian conjecture. In *Annales Polonici Mathematici*, volume 110, pages 1–11. Institute of Mathematics Polish Academy of Sciences, 2014.

P. Chakraborty, M. M. Meerschaert, and C. Y. Lim. Parameter estimation for fractional transport: A particle-tracking approach. *Water resources research*, 45(10), 2009.

D. Chen, R. Carsel, L. Moeti, and B. Vona. Assessment and prediction of contaminant transport and migration at a florida superfund site. *Environmental monitoring and assessment*, 57:291–299, 1999.

T. Chen, G.-H. Liu, and E. A. Theodorou. Likelihood training of schr\" odinger bridge using forward-backward sdes theory. *arXiv preprint arXiv:2110.11291*, 2021.

L. Chizat, S. Zhang, M. Heitz, and G. Schiebinger. Trajectory inference via mean-field langevin in path space. *Advances in Neural Information Processing Systems*, 35:16731–16742, 2022.

M. A. Coomer, L. Ham, and M. P. Stumpf. Noise distorts the epigenetic landscape and shapes cell-fate decisions. *Cell Systems*, 13(1):83–102, 2022.

T. M. Cover and J. A. Thomas. *Elements of information theory*, volume 2. Wiley New York, 1991.

V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

J. du Buisson and H. Touchette. Dynamical large deviations of linear diffusions. *Physical Review E*, 107(5):054111, 2023.

P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In *International conference on machine learning*, pages 1367–1376. PMLR, 2018.

A. Einstein. *Investigations on the Theory of the Brownian Movement*. Courier Corporation, 1956.

A. M. Elfeki. Prediction of contaminant plumes (shapes, spatial moments and macrodispersion) in aquifers with insufficient geological information. *Journal of Hydraulic Research*, 44(6):841–856, 2006.

W. Fang and M. B. Giles. Adaptive euler–maruyama method for sdes with nonglobally lipschitz drift. *The Annals of Applied Probability*, 30(2):526–560, 2020.

J. A. Farrell, Y. Wang, S. J. Riesenfeld, K. Shekhar, A. Regev, and A. F. Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360 (6392):eaar3131, 2018.

A. Forrow. Consistent diffusion matrix estimation from population time series. *arXiv preprint arXiv:2408.14408*, 2024.

E. Frind, J. Molson, and D. Rudolph. Well vulnerability: a quantitative approach for source water protection. *Groundwater*, 44(5):732–742, 2006.

Z. Gajic and M. T. J. Qureshi. *Lyapunov matrix equation in system stability and control*. Courier Corporation, 2008.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 1574–1583. PMLR, 2019.

P. Ghosal, M. Nutz, and E. Bernton. Stability of entropic optimal transport and schrödinger bridges. *Journal of Functional Analysis*, 283(9):109622, 2022.

V. Guan, J. Janssen, N. Lanzetti, A. Terpin, G. Schiebinger, and E. Robeva. Gradient-flow sdes have unique transient population dynamics. *arXiv preprint arXiv:2505.21770*, 2025.

N. Hansen and A. Sokol. Causal interpretation of stochastic differential equations. *Electronic Journal of Probability*, 2014.

T. Hashimoto, D. Gifford, and T. Jaakkola. Learning population-level diffusions with generative rnns. In *International Conference on Machine Learning*, pages 2417–2426. PMLR, 2016.

M. B. Hazas, F. Ziliotto, J. Lee, M. Rolle, and G. Chiogna. Evolution of plume geometry, dilution and reactive mixing in porous media under highly transient flow fields at the surface water-groundwater interface. *Journal of Contaminant Hydrology*, 258:104243, 2023.

H. Huang. One-step data-driven generative model via schrödinger bridge. *arXiv preprint arXiv:2405.12453*, 2024.

H. Janati. *Advances in Optimal transport and applications to neuroscience*. PhD thesis, Institut Polytechnique de Paris, 2021.

H. Janati, B. Muzellec, G. Peyré, and M. Cuturi. Entropic optimal transport between unbalanced gaussian measures has a closed form. *Advances in neural information processing systems*, 33:10468–10479, 2020.

S. Karlin and W. J. Studden. *Tchebycheff systems: With applications in analysis and statistics*. Interscience Publishers, 1966.

H. Lavenant, S. Zhang, Y.-H. Kim, and G. Schiebinger. Towards a mathematical theory of trajectory inference. *arXiv preprint arXiv:2102.09204*, 2021.

J. Ledolter. Estimation bias in the first-order autoregressive model and its impact on predictions and prediction intervals. *Communications in Statistics-Simulation and Computation*, 38(4):771–787, 2009.

P. C. Lichtner, S. Kelkar, and B. Robinson. New form of dispersion tensor for axisymmetric porous media with implementation in particle tracking. *Water Resources Research*, 38 (8):21–1, 2002.

L. Locatelli, P. J. Binning, X. Sanchez-Vila, G. L. Søndergaard, L. Rosenberg, and P. L. Bjerg. A simple contaminant fate and transport modelling tool for management and risk assessment of groundwater pollution from contaminated sites. *Journal of contaminant hydrology*, 221:35–49, 2019.

L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3):218–229, 2021.

J. MacInnes and F. Bracco. Stochastic particle dispersion modeling and the tracer-particle limit. *Physics of Fluids A: Fluid Dynamics*, 4(12):2809–2824, 1992.

D. Mackay, D. Freyberg, P. Roberts, and J. Cherry. A natural gradient experiment on solute transport in a sand aquifer: 1. approach and overview of plume movement. *Water Resources Research*, 22(13):2017–2029, 1986.

S. Maddu, V. Chardès, M. Shelley, et al. Inferring biological processes with intrinsic noise from cross-sectional data. *arXiv preprint arXiv:2410.07501*, 2024.

C. Man and C. W. Tsai. Stochastic partial differential equation-based model for suspended sediment transport in surface water flows. *Journal of engineering mechanics*, 133(4): 422–430, 2007.

S. Mandt, M. Hoffman, and D. Blei. A variational analysis of stochastic gradient algorithms. In *International conference on machine learning*, pages 354–363. PMLR, 2016.

G. Manten, C. Casolo, E. Ferrucci, S. W. Mogensen, C. Salvi, and N. Kilbertus. Signature kernel conditional independence tests in causal discovery for stochastic processes. *arXiv preprint arXiv:2402.18477*, 2024.

J. Marcinkiewicz. Sur une propriété de la loi de gauss. *Mathematische Zeitschrift*, 44(1): 612–618, 1939.

C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

B. Mityagin. The zero set of a real analytic function. *arXiv preprint arXiv:1512.07276*, 2015.

S. W. Mogensen and N. R. Hansen. Graphical modeling of stochastic processes driven by correlated noise. *Bernoulli*, 28(4):3023–3050, 2022.

K. Neklyudov, R. Bondesan, and M. Welling. Deterministic gibbs sampling via ordinary differential equations. *arXiv preprint arXiv:2106.10188*, 2021.

V. Nenna, A. Pidlisecky, and R. Knight. Application of an extended kalman filter approach to inversion of time-lapse electrical resistivity imaging data for monitoring recharge. *Water Resources Research*, 47(10), 2011.

J. N. Nielsen, H. Madsen, and P. C. Young. Parameter estimation in stochastic differential equations: an overview. *Annual Reviews in Control*, 24:83–94, 2000.

M. Nutz. Introduction to entropic optimal transport. *Lecture notes, Columbia University*, 2021.

D. R. O'Connor et al. Part two: Report of the walkerton inquiry: A strategy for safe drinking water. Technical report, Ministry of the Attorney General, 2002.

W. Ogle et al. *Aristotle: on the parts of animals*. Kegan Paul, French & Company, 1882.

J. Oh and C. W. Tsai. A stochastic jump diffusion particle-tracking model (sjd-ptm) for sediment transport in open channel flows. *Water Resources Research*, 46(10), 2010.

B. Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

A. J. Paulson. The transport and fate of fe, mn, cu, zn, cd, pb and so4 in a groundwater plume and in downstream surface waters in the coeur d'alene mining district, idaho, usa. *Applied Geochemistry*, 12(4):447–464, 1997.

G. A. Pavliotis. Stochastic processes and applications. *Texts in Applied Mathematics*, 60, 2014.

R. Pawlowicz, C. Hannah, and A. Rosenberger. Lagrangian observations of estuarine residence times, dispersion, and trapping in the salish sea. *Estuarine, Coastal and Shelf Science*, 225:106246, 2019.

S. Peiris, R. Mellor, and P. Ainkaran. Maximum likelihood estimation for short time series with replicated observations: A simulation study. *InterStat*, 9:1–16, 2003.

K. B. Petersen, M. S. Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

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

A. Reisach, C. Seiler, and S. Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

M. Rohbeck, B. Clarke, K. Mikulik, A. Pettet, O. Stegle, and K. Ueltzhöffer. Bicycle: Intervention-based causal discovery with cycles. In *Causal Learning and Reasoning*, pages 209–242. PMLR, 2024.

J. Runge. Necessary and sufficient graphical conditions for optimal adjustment sets in causal graphical models with hidden variables. *Advances in Neural Information Processing Systems*, 34:15762–15773, 2021.

P. Salamon, D. Fernàndez-Garcia, and J. Gómez-Hernández. Modeling tracer transport at the made site: The importance of heterogeneity. *Water Resources Research*, 43(8), 2007.

A. Santos, D. Rente, R. Seabra, and J. M. Moura. Learning the causal structure of networked dynamical systems under latent nodes and structured noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14866–14874, 2024.

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

G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

M. Setty, M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe'er. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature biotechnology*, 34(6):637–645, 2016.

Y. Shen, R. Berlinghieri, and T. Broderick. Learning a vector field from snapshots of unidentified particles rather than particle trajectories. In *ICLR 2024 Workshop on AI4DifferentialEquations In Science*, 2024.

Y. Shen, R. Berlinghieri, and T. Broderick. Multi-marginal schrödinger bridges with iterative reference refinement. In Y. Li, S. Mandt, S. Agrawal, and E. Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3817–3825. PMLR, 03–05 May 2025. URL `https://proceedings.mlr.press/v258/shen25b.html`.

Y. Shi, V. De Bortoli, A. Campbell, and A. Doucet. Diffusion schrödinger bridge matching. *Advances in Neural Information Processing Systems*, 36, 2024.

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

A. Tejada-Lapuerta, P. Bertin, S. Bauer, H. Aliee, Y. Bengio, and F. J. Theis. Causal machine learning for single-cell genomics. *arXiv preprint arXiv:2310.14935*, 2023.

A. Terpin, N. Lanzetti, and F. Dörfler. Learning diffusion at lightspeed. *arXiv preprint arXiv:2406.12616*, 2024.

C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4): 381–386, 2014.

F. Vargas, P. Thodoroff, A. Lamacraft, and N. Lawrence. Solving schrödinger bridges via maximum likelihood. *Entropy*, 23(9):1134, 2021.

C. Waddington. *How Animals Develop*. George Allen & Unwin, 1935.

Y. Wang, X. Geng, W. Huang, B. Huang, and M. Gong. Generator identification for linear sdes with additive and multiplicative noise. *Advances in Neural Information Processing Systems*, 36, 2024.

C. Weinreb, S. Wolock, B. K. Tusi, M. Socolovsky, and A. M. Klein. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences*, 115(10):E2467–E2476, 2018.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, pages 1–25, 1982.

R. E. Williamson and H. F. Trotter. *Multivariable mathematics: linear algebra, differential equations, calculus*. Prentice-Hall, 1974.

T. Yachimura, H. Wang, Y. Imoto, M. Yoshida, S. Tasaki, Y. Kojima, Y. Yabuta, M. Saitou, and Y. Hiraoka. scegot: Single-cell trajectory inference framework based on entropic gaussian mixture optimal transport. *bioRxiv*, pages 2023–09, 2023.

S. Y. Zhang. Joint trajectory and network inference via reference fitting. *arXiv preprint arXiv:2409.06879*, 2024.

W. Zhao, E. Larschan, B. Sandstede, and R. Singh. Optimal transport reveals dynamic gene regulatory networks via gene velocity estimation. *bioRxiv*, pages 2024–09, 2024.