Identifying drift, diffusion, and causal structure from temporal snapshots

Vincent Guan[1] Joseph Janssen[1] Hossein Rahmani[1] Andrew Warren[1]
Stephen Zhang[2] Elina Robeva[1] Geoffrey Schiebinger[1]

[1]University of British Columbia
[2]University of Melbourne

October 16, 2024

Abstract

Stochastic differential equations (SDEs) fundamentally characterize dynamic processes, including gene regulatory networks (GRNs), contaminant transport, financial markets, and image generation in computer vision. However, learning the underlying stochastic model from observational data is a challenging task, especially in settings where observations are restricted to snapshots of the process at certain measurement times. Motivated by burgeoning snapshot-based single-cell datasets, there has been a flurry of work on algorithms for trajectory and gene regulatory network (GRN) inference. In this work, we first provide a complete characterization of system identifiability under the linear additive noise model. In particular, the drift and diffusion parameters can be guaranteed to be identified from temporal marginals, if and only if the initial marginal is not auto-rotationally invariant. Furthermore, we show that the underlying causal graph of any additive noise system can be recovered from the drift and diffusion. We then generalize the entropic optimal transport problem to introduce Alternating Projection Parameter Estimation from X_0 (APPEX), a novel method for identifying both drift and arbitrary additive diffusion, without prior knowledge. We prove that each step in the APPEX algorithm is optimal with respect to the Kullback–Leibler divergence, and we conduct several experiments on simulated data to demonstrate its effectiveness.

1 Introduction

This work presents the first comprehensive approach for simultaneously estimating the drift and diffusion of a stochastic differential equation (SDE) from observed temporal marginals. Parameter estimation has been studied extensively from trajectory data, either given one long trajectory [BPRF06, NMY00, Bis07, TB13, MMP15, KOLL12, NR20, CHLS23], or multiple short trajectories [MCF+24, WGH+24, LJP+21, PHR19, NM]. However, individual trajectories may not always be observable in many applications

In many applications, we may only observe snapshots of the process at various times. This setting is prominent in single-cell RNA (scRNA) datasets, where the goal is to learn how gene concentrations evolve in cells, but sequencing measurement technologies are destructive [SST+19, LZKS21]. Similarly, we may only observe temporal marginals when studying contaminant flow, where the goal is to learn the hydrological dynamics governing plume migration, since hydrogeochemical sensors cannot track individual particles [MT07, HZL+23, SFGGH07, Elf06, AG92, BYB+92, MFRC86].

Since the drift and diffusion determine trajectory fates and the system's causal structure, parameter estimation is a highly motivated area of research, particularly for trajectory inference [SST+19, WWT+18, LZKS21, CZHS22, YWI+23, SBB24, Zha24] and gene regulatory network inference [AVI+20, ZLSS24, ATW+24, RCM+24, Zha24] from scRNA datasets. However, due to challenges in stochastic analysis posed by observing only temporal marginals, namely sources of non-identifiability, previous works assume that at least one of diffusion or drift are already known, in order to perform the inference.

1.1 Background and related work

SDEs have long been used to model natural processes, such as contaminant transport [OT10, DAD05], population dynamics [LES03], and gene expression trajectories [SST+19]. The drift and diffusion often offer

physical interpretations. In scRNA datasets, the drift identifies the set of genes that directly regulate the expression of a gene of interest [Zha24, AVI+20, ATW+24, ZLSS24, TLBB+23]. Drift estimation therefore provides important insights into studying genetic diseases and devising gene therapies. Estimating diffusion is also important because the magnitude of diffusion influences the fate of cells moving to various steady-states or wells [For24]. While small diffusion implies that cell fates are mostly determined by their initial conditions, large-scale diffusion implies that cell fates are significantly influenced by random or unknown processes [For24]. Likewise, understanding pollutant fates is a central problem in drinking water protection [O+02, FMR06, LBSV+19, CCMV99, Pau97, CBL+19]. In hydrological systems, drift is linked to important and unknown properties of the subsurface, such as hydraulic conductivity through average flow velocity [HZL+23, BHR93]. Diffusion in contaminant transport problems usually describes material heterogeneity or turbulence [BHR93], and therefore provides insight into the applicability of popular models, such as Darcy's law [MT07, OT10, LKR02]. While a particular application may only call for the estimation one or the other, drift and diffusion are independently important and inextricably linked. As we will show in this paper, a good estimate of one can only be obtained given a good estimate of the other. Drift and diffusion also can be linked more directly in the examples of the chemical Langevin equation and the advection-dispersion equation for solute transport [HZL+23, LKR02, BHR93], where drift and diffusion share the same parameters [Gil00].

Inference from temporal marginals. Within the marginals-only observational setting, many works focus on trajectory inference [SST+19, LZKS21, YWI+23, CZHS22], while others infer the causal graph (e.g. GRN) [AVI+20, BLL+20, TLBB+23, RCM+24] or perform parameter estimation [CML09]. However, since these quantities are fundamentally interconnected, recent research leverages these relationships to jointly estimate a subset of these quantities, often with one primary objective. For instance, [VTLL21, SBB24] iteratively estimate drift and trajectories, for trajectory inference. Meanwhile, [Zha24] iteratively estimates drift and trajectories, while additionally applying permanent interventions on the drift dynamics for network inference.

Importantly, previous works assume that a key quantity is already known, prior to inference. Most approaches assume that the underlying diffusion is known [LZKS21, SBB24, VTLL21, CZHS22, Zha24], or derived from a possibly misspecified reference SDE [OT10]. In contrast, [For24] estimates diffusion from temporal marginals, but assumes that an accurate estimate of drift is known. However, knowing drift or diffusion *a priori* reduces the inference problem such that only one component of the underlying SDE is unknown, which is unrealistic in practice. Due to their intrinsic relationship in generating the data, misspecified diffusion often leads to poor drift estimation and vice-versa. In addition to assuming prior knowledge, previous works also impose model simplifications. The standard assumption is that diffusion is modeled by isotropic Brownian motion σdW_t [LZKS21, CZHS22, Zha24, SBB24, VTLL21, WWT+18]. However, this prevents different noise scales across variables and correlated noise across variables from latent confounders [RCM+24, MH22]. Similarly, it is common to assume that drift is irrotational [WWT+18, LZKS21, CZHS22], or small/sparse [Zha24] to avoid nonidentifiability issues. However, rotational drift corresponds to cycles in the causal graph, which are important features of GRNs. In contrast to previous related work, we jointly estimate the system's SDE parameters, causal graph, and trajectories, solely from the temporal marginals, without imposing irrotational drift or isotropic diffusion.

Identifiability. Previous assumptions on knowing diffusion or drift a priori are largely attributed to sources of non-identifiability introduced by the marginals-only observational setting [WWT+18, CHS22, LZKS21, For24]. Even with these assumptions, it is common to leverage additional perturbational data from interventions, to resolve cycles and non-identifiability [RCM+24, Zha24]. Establishing sufficient conditions for identifiability are therefore required to ensure that inference is feasible. [WGH+24] 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, and the SDE parameters, A and GG^T. In contrast, our work studies identifiability of linear additive noise SDEs in the more general setting of observing temporal marginals. As detailed in Section 3, the set of non-identifiable systems within our observational setting is strictly larger. Furthermore, we demonstrate that identifiability can be guaranteed with a wide class of initial distributions, regardless of the linear additive noise SDE. To the best of our knowledge, this has not been shown even in the case of continuous trajectory observations.

By connecting SDE parameter identification to recent work on dynamic structural causal models [BM24], we uncover interesting parallels to the well-studied linear non-Gaussian graphical model from the static causal setting, under which full causal identification is possible [SHH+06]. In particular, we show that full causal identification is possible from the linear additive noise SDE model, if the initial distribution X_0 is not

auto-rotationally invariant (which implies non-Gaussianity).

1.2 Our contributions

In this work, we show that it is possible to infer the drift, diffusion, trajectories, and the underlying causal graph, solely from temporal marginals. We primarily consider linear additive noise models, which are the most studied SDE [WGH+24, YWI+23, Löb14, Zha24, RCM+24]. We derive theoretical conditions that make this inference possible, and develop a method to perform the joint inference. Our contributions are summarized in Figure 1.

1. We provide a theoretical foundation for system identification from temporal marginals in Section 4.
   (a) We provide a full characterization of the identifiability conditions for the drift and diffusion parameters from temporal marginals, under the linear additive noise model. Identifiability is guaranteed if and only if the initial distribution X_0 is not auto-rotationally invariant.
   (b) We connect parameter estimation to causal structure learning via dynamic structural causal models [BM24]. We show that the causal graph of any additive noise model can be recovered from the SDE parameters.
2. In Section 5, we introduce the Alternating Projection Parameter Estimation from X_0(APPEX) algorithm—the first method designed to estimate drift and diffusion from temporal marginals without relying on prior knowledge. APPEX is also formulated for general additive noise GdW_t, extending beyond the standard assumption of isotropic noise σdW_t.
   (a) We show that with each iteration, APPEX's estimates approach the true solution. In particular, APPEX alternates between a trajectory inference step, which is a Schrodinger bridge problem, and a parameter estimation step, which is solved via maximum likelihood estimation. We prove that both subprocedures are optimal with respect to KL divergence.
   (b) To solve the associated Schrodinger bridge problem for trajectory inference, given arbitrary additive noise, we provide a generalization of the entropic optimal transport problem for non-scalar entropic regularization.

We test APPEX's efficacy across a wide range of experiments in Section 6. We first show that APPEX can identify the drift and diffusion parameters from the previously non-identifiable examples introduced in Section 3, given appropriate non-auto-rotationally invariant initial distributions. We also 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 [SST+19]. Our final experiments demonstrate that APPEX can be used to identify the system's causal graph, even in the presence of latent confounders. Although in experimental settings APPEX's estimates seem to approach the true solution and seem not to depend on our initial guesses for A and H, we have not yet proved convergence, due to the non-convexity of the optimization space.

2 Mathematical setup

Let $X_t \in \mathbb{R}^d$, $G \in \mathbb{R}^{d\times m}$, and $W_t \in \mathbb{R}^m$ be m dimensional Brownian motion. An additive noise SDE obeys the form

$$dX_t = b(X_t)dt + GdW_t, \quad X_0 \sim p_0, \tag{1}$$

where $b(X_t)$ is a Lipschitz drift function $b : \mathbb{R}^d \to \mathbb{R}^d$, to guarantee existence and uniqueness of the SDE solution [HC15], and $G \in \mathbb{R}^{d\times m}$ is the diffusion matrix. In this paper, we will also refer to $H = GG^T \in \mathbb{R}^{d\times d}$ as the (observational) diffusion, since it is only possible to observe H [WGH+24, Pav14].

An SDE defines a law on paths, which governs the probablistic evolution of individual trajectories. Indeed, if the process has bounded second moments, and is defined by a Lipschitz drift function and nondegenerate diffusion, then the process admits a density with respect to the Wiener measure over the path space $\Omega = C([0,T],\mathbb{R}^d)$ [Oks13, SBB24]. Moreover, trajectories evolve according to the transition kernel, which can be approximated for additive noise SDEs with $X_{t+dt}|X_t \sim \mathcal{N}(X_t + b(X_t)dt, Hdt)$ [SBB24, Pav14].

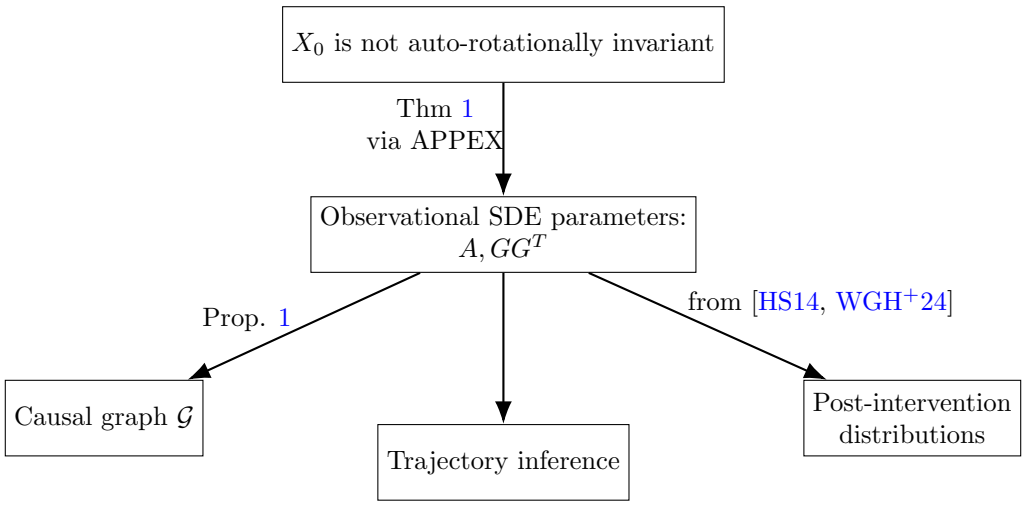


Figure 1: A good initial marginal (and our algorithm APPEX) is all you need. Under a linear additive noise model, one can recover the drift, diffusion, causal graph, trajectories, and post-intervention distributions from temporal marginals, provided that X_0 is not auto-rotationally invariant.

Linear additive noise SDEs are the most commonly studied additive noise SDE, and have been applied to numerous fields of interest [WGH+24]. These SDEs generalize the popular Ornstein-Uhlenbeck process, such that stationarity is not assumed. A linear additive noise SDE obeys the form

$$dX_t = AX_t dt + GdW_t, \quad X_0 \sim p_0, \tag{2}$$

where $A \in \mathbb{R}^{d\times d}$. The exact transition kernel is given by $X_{t+dt}|X_t \sim \mathcal{N}(\mu_t, \Sigma_t)$, where $\mu_t = e^{Adt}X_t$ and $\Sigma_t = \int_t^{t+dt} e^{A(t-s+dt)} H e^{A^T(t-s+dt)}\, ds$ [Zha24], and can practically be estimated via $\mathcal{N}(X_t + AX_t dt, Hdt)$.

The goal of SDE parameter estimation under the most general observational setting is to infer the drift and diffusion parameters from observed temporal marginals, or "snapshots", of the process. Formally, we consider a set of N observed marginals $\{p_i\}_{i=0}^{N-1}$ over times $t_0, \dots t_{N-1}$, where the ith marginal is

$$p_i \sim unif\{x_{t_i}^{(j)} : j \in 1, \dots M_i\},$$

such that $x_{t_i}^{(j)}$ is the jth observation of the process X_t at time t_i, and M_i is the number of observations at time t_i.For linear additive noise SDEs, the goal is to identify A and $H = GG^T$ from the marginals. if we take the number of observations M_i per time to be infinite, then our observed marginals would be given by the marginal distributions of the SDE

$$p_i \sim X_{t_i}, p_0 \sim X_0.$$

The Fokker-Planck equation (3) describes the evolution of the process' temporal marginals from the initial distribution p_0. A straightforward computation shows that the marginals of a linear additive noise SDE evolve according to the following version of the Fokker-Planck equation:

$$\frac{\partial}{\partial_t} p(x,t) = -\nabla \cdot (Ax)p(x,t) + \frac{H}{2}\nabla^2 p(x,t), \quad p(x,0) = p_0. \tag{3}$$

Although (3) assumes differentiability, we note that it can also be defined in the weak distributional sense (see Appendix Section A.2).

3 Examples of non-identifiability

The problem of non-identifiability given p_0 appears if there exists an alternative drift-diffusion pair, $(A_1, H_1) \neq (A, H)$, which shares the same time marginals, following initial distribution $X_0 \sim p_0$. Equivalently, the processes share the same Fokker-Planck equation (3) across all observed times. We first summarize a few classical examples of non-identifiable SDE pairs from the literature [LZKS21, SBB24, WGH+24, HGJ16, WWT+18].

Pair 1: Starting at stationary distribution [LZKS21, SBB24]

$$dX_t = -X_t dt + dW_t, X_0 \sim \mathcal{N}(0, \frac{1}{2}) \tag{4}$$

$$dY_t = -10Y_t dt + \sqrt{10} dW_t, Y_0 \sim \mathcal{N}(0, \frac{1}{2}) \tag{5}$$

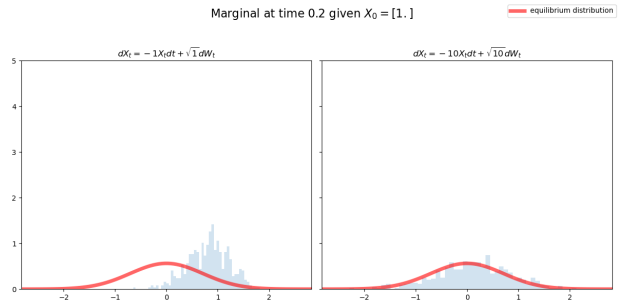
In this example, both SDEs have the same stationary distribution $\mathcal{N}(0, \frac{1}{2})$ despite having significantly different individual trajectories. Indeed, the stationary distribution of a 1-dimensional 0-mean Ornstein-Uhlenbeck (OU) process with drift $-\lambda X_t$ and diffusion σ is $\mathcal{N}(0, \frac{\sigma^2}{2d\lambda})$, which depends only on the drift:diffusivity ratio λ/σ^2 [LZKS21]. Hence, if $p_0 \sim \mathcal{N}(0, \frac{1}{2})$, then X_t and Y_t are non-identifiable from one another when only observing the marginals. For multivariate OU processes with drift A and observational diffusion H, the stationary distribution $\mathcal{N}(0, \Sigma)$ depends only on the relationship $\Sigma A + A\Sigma = -H$ [MHB16].

Pair 2: Rotation around process mean [SBB24, HGJ16, WWT+18]

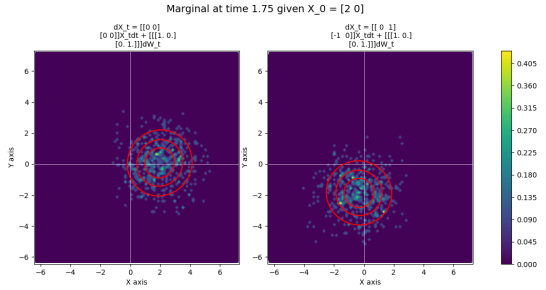
$$dX_t = dW_t, X_0 \sim \mathcal{N}(0, Id) \tag{6}$$

$$dY_t = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} Y_t dt + dW_t, , Y_0 \sim \mathcal{N}(0, Id) \tag{7}$$

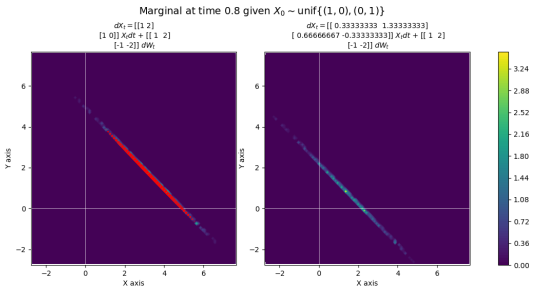
In this example, the first SDE governing X_t is 2-dimensional Brownian motion and the second SDE governing Y_t adds a divergence-free rotational vector field $(x, y) \to (y, -x)$ about $(0, 0)$. Hence, if p_0 is an isotropic distribution with mean $(0, 0)$, then X_t and Y_t are non-identifiable from one another [SBB24]. This can also be shown directly with the Fokker-Planck equation (3), since $\nabla \cdot (Ax)p(x,t) = \nabla p(x,t) \cdot Ax + p(x,t)\nabla \cdot A(x) = \nabla p(x,t) \cdot Ax = 0$, if p is parallel to the rotational vector field $Ax = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} x$.



(a) Example (5), $p_0 = \delta(x - 1)$. The SDE with $\lambda = \sigma^2 = 10$ (right) has converged to the stationary distribution at time $t = 0.2$, while the SDE with $\lambda = \sigma^2 = 1$ (left) has not yet sufficiently drifted or diffused.



(b) Example (7), $p_0 = \delta(x - (2, 0)^T)$. The first SDE (left) maintains mean $(2, 0)$. The second SDE (right) rotates about the circle $x^2 + y^2 = 4$. At time $t = 1.75$, the mean has shifted to approximately $(0, -2)$.



(c) Example (9), $p_0 \sim unif\{(1, 0)^T, (0, 1)^T\}$. 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 2: Marginals are plotted at various times for the SDE pairs from Section 4 following our proposed initialization.

Pair 3: Degenerate rank [WGH+24]

$$dX_t = \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix} X_t dt + \begin{bmatrix} 1 & 2 \\ -1 & -2 \end{bmatrix} dW_t, X_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{8}$$

$$dY_t = \begin{bmatrix} 1/3 & 4/3 \\ 2/3 & -1/3 \end{bmatrix} Y_t dt + \begin{bmatrix} 1 & 2 \\ -1 & -2 \end{bmatrix} dW_t, Y_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{9}$$

In this example motivated by [WGH+24], 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 are non-identifiable from one another. We note that this non-identifiability holds even when we observe trajectories [WGH+24], whereas the first two examples are identifiable from trajectories but not identifiable from marginals.

4 System identification from temporal marginals

In this section, we begin with Section 4.1 where we provide a full characterization of the conditions for identifiability of linear additive noise SDEs from observed time marginals. In particular, we will see that identifiability is determined by generalized rotational properties of the initial distribution X_0, and that one can recover the causal graph from the identified parameters. We then show in Section 4.2 that identifying an additive noise SDE also allows us to recover the underlying causal graph.

Definition 1. *We define a Σ-generalized rotation in $\mathbb{R}^d$ as the matrix exponential e^{At}, such that $A \in \mathbb{R}^{d\times d}$ is skew-symmetric with respect to $\Sigma \succeq 0$, i.e. $A\Sigma + \Sigma A^T = 0$.*

We note that $\Sigma = Id$ defines classical rotations, such that spheres, e.g. $S^{d-1} = \{x \in \mathbb{R}^d : x^T x = 1\}$, are rotationally invariant. For non-isotropic choices of Σ, ellipsoids, e.g. $E_\Sigma = \{x \in \mathbb{R}^d : x^T\Sigma^{-1}x = 1\}$, are rotationally invariant. If $\Sigma \succ 0$, then Σ-generalized rotations can be interpreted as classical rotations within the Σ-weighted inner product space $\mathbb{R}^d_\Sigma$, since they preserve the norm $\|x\|^2_{\Sigma,2} = x^T\Sigma^{-1}x$.

Definition 2. *Let X be a d-dimensional r.v. with covariance Σ. We define X to be auto-rotationally invariant, if there exists a nontrivial $A \neq 0$ s.t. $e^{At}X \sim X\ \forall t \geq 0$. Conversely, X is not auto-rotationally invariant, if $e^{At}X \sim X\ \forall t \geq 0$ admits only the solution $A = 0$.*

We note that the condition $e^{At}X \sim X\ \forall t \geq 0$ implies that the only feasible solutions A are skew-symmetric with respect to Σ (Lemma 1 in the Appendix). Auto-rotational invariance therefore relates directly to Σ-generalized rotations, and can be interpreted as a random variable being rotationally invariant under the geometry induced by its own covariance Σ. For example, Gaussians $\mathcal{N}(0, \Sigma)$ and uniform distributions over ellipsoids are auto-rotationally invariant random variables. Additionally, all rank degenerate r.v.'s are auto-rotationally invariant, since $e^{At}X \sim X$ would be satisfied by any matrix A, such that X is in its nullspace with probability 1. However, given full rank conditions, auto-rotational invariance requires strict ellipsoidal symmetry on the r.v.'s density (Lemma 2 in Appendix), which means that almost all non-degenerate r.v's are not auto-rotationally invariant. We provide more examples and theory about rotations in Section A.3 of the Appendix, and also refer the reader to [Özd16] for additional theory about elliptical rotations.

4.1 SDE parameter identifiability

We are now ready for our main identifiability theorem, which completely characterizes identifiability of linear additive noise SDEs using auto-rotational invariance.

Theorem 1. *Let X_t evolve according to a d-dimensional linear additive noise SDE (2), with initial condition $X_0 \sim p_0$, such that all the moments of p_0 are finite. Then, the drift A and the diffusion GG^T of the SDE can be guaranteed to be uniquely identified from the time marginals p_t if and only if $X_0 \sim p_0$ is not auto-rotationally invariant.*

Proof. We first prove that if X_0 is auto-rotationally invariant, then there exist multiple processes with distinct drift-diffusion parameters $(A, H = GG^T)$, which would share the same time marginals p_t, when initialized at $X_0 \sim p_0$. The idea is to generalize the non-identifiable isotropic rotation example (7). We first reason in the case where p_t has smooth density $p(x,t)$, then afterwards consider the case where p_t is any probability measure (which follows by an analogous but more abstract argument).

Let X_0 be an auto-rotationally invariant r.v. with covariance $\Sigma = GG^T$. Then, there exists $A \neq 0$ such that $e^{As}X_0 \sim X_0 \ \forall s \geq 0$. In particular, we will show that for all $\gamma \in \mathbb{R}$, the SDEs

$$dX_t = \gamma G dW_t \tag{10}$$
$$dX_t = AX_t dt + \gamma G dW_t \tag{11}$$

will have the same time marginals $p(x,t)$. The Fokker-Planck equations for the two SDEs are given by

$$\frac{\partial}{\partial t}p(x,t) = \frac{\gamma^2\Sigma}{2}\nabla^2 p(x,t) \qquad p(x,0) \sim p_0, \tag{12}$$
$$\frac{\partial}{\partial t}p(x,t) = -\nabla \cdot (Ax)p(x,t) + \frac{\gamma^2\Sigma}{2}\nabla^2 p(x,t) \qquad p(x,0) \sim p_0. \tag{13}$$

To show that the processes exhibit the same time marginals, we show that (12) and (13) are equivalent by computing that the divergence term in (13) is identically zero. We first prove equivalence at $t = 0$, by showing that for all $x \in \mathbb{R}^d$,

$$\nabla \cdot (Ax)p(x,0) = \nabla p(x,0) \cdot Ax + p(x,0)\nabla \cdot (Ax) = 0. \tag{14}$$

By the auto-rotational invariance of X_0, $p(x,0)$ satisfies: $p(e^{As}x, 0) = p(x,0) \quad \forall s$. Taking the derivative with respect to s at $s = 0$ and using chain rule, we have:

$$0 = \frac{d}{ds}p(x,0)\Big|_{s=0} = \frac{d}{ds}p(e^{As}x,0)\Big|_{s=0} = \nabla p(x,0) \cdot \frac{d}{ds}e^{As}x\Big|_{s=0} = \nabla p(x,0) \cdot Ax,$$

so the first term on the right side of (14) vanishes. We now show $\nabla \cdot (Ax) = \mathrm{Tr}(A) = 0$ to prove (14). Lemma 1 implies that $A\Sigma + \Sigma A^T = 0$ and Lemma 5 shows that without loss of generality, we may pick A so that $Tr(A) = 0$. This proves (14) for $t = 0$. The same argument applies for $t > 0$ upon noticing that for both systems, X_t is also auto-rotationally invariant with respect to the same map $e^{As}X_t \sim X_t$. The details are given in Lemma 6. Hence, if X_0 is auto-rotationally invariant, then we do not have identifiability from marginals over the class of linear additive noise SDEs.

Moreover the same reasoning applies even in the case where p_t does not have density, since we can instead work with the weak formulation of the Fokker-Planck equation. It will suffice to show that for all $\varphi \in C_c^2(\mathbb{R}^d)$, we have that

$$\int (\nabla\varphi(x) \cdot Ax)dp_t(x) = 0$$

since this corresponds to the weak formulation of the divergence term $\nabla \cdot (p_t(x)Ax)$. Integrating by parts, we see that

$$\int (\nabla\varphi(x) \cdot Ax)dp_t(x) = -\int \varphi(x)\nabla \cdot (Ax)dp_t(x).$$

We have already established that $\nabla \cdot (Ax) = 0$, hence $\varphi(x)\nabla \cdot (Ax) = 0$ also. The claim follows.

We now prove that identifiability is guaranteed if X_0 is not auto-rotationally invariant. Suppose that two linear additive noise SDEs with parameters $(A, H = GG^T)$ and $(A_1, H_1 = G_1G_1^T)$ have the same time marginals $p(x,t)$ given the same initial distribution $X_0 \sim p_0$, which is not auto-rotationally invariant. We will prove that $(A, H) = (A_1, H_1)$ must hold. Since both SDEs share the same marginals, their Fokker-Planck equations are equivalent:

$$\frac{\partial}{\partial_t}p(x,t) = L(p(x,t)) := -\nabla \cdot (Ax)p(x,t) + \frac{H}{2}\nabla^2 p(x,t) \qquad p(x,0) \sim p_0 \tag{15}$$
$$= L_1(p(x,t)) := -\nabla \cdot (A_1x)p(x,t) + \frac{H_1}{2}\nabla^2 p(x,t) \qquad p(x,0) \sim p_0 \tag{16}$$

By linearity, we may subtract both equations and obtain

$$0 = (L - L_1)(p(x,t)) = -\nabla \cdot (\bar{A}x)p(x,t) + \frac{\bar{H}}{2}\nabla^2 p(x,t) \qquad p(x,0) \sim p_0, \tag{17}$$

where $\bar{A} = A - A_1$ and $\bar{H} = H - H_1$. We note that this result also holds if we consider the weak distributional sense (A.2) of the Fokker-Planck equation. Indeed, let us compute a version of (16) for the weak sense of the Fokker-Planck equation. For each t, let p_t denote the probability measure which is the distribution of X_t. For any test function $\varphi \in C_c^2(\mathbb{R}^d)$,

$$\frac{d}{dt}\int \varphi(x)dp_t(x) = \int L^*\varphi(x)dp_t(x) = \int L_1^*\varphi(x)dp_t(x)$$

which implies that for all t

$$0 = \int (L - L_1)^*\varphi(x)dp_t(x). \tag{18}$$

Now we take $L - L_1$ to be the differential operator for the residual SDE. From the uniqueness theory for the weak solutions of the Kolmogorov forward equation (see [Str08] Thm. 2.2.9) it follows that as long as $p(x, 0)$ has finite moments of all orders, then there is exactly one weakly continuous solution $\mu_t : [0, T] \to \mathcal{P}(\mathbb{R}^d)$ beginning at p_0 satisfying, for all $\varphi \in C_c^2(\mathbb{R}^d)$

$$\frac{d}{dt}\int \varphi(x)d\mu_t(x) = \int (L - L_1)^*\varphi(x)d\mu_t(x).$$

This implies that for the initial condition p_0, we have in particular that $0 = \int (L - L_1)^*\varphi(x)dp_0(x)$, which implies that $\frac{d}{dt}\int \varphi(x)d\mu_t(x)|_{t=0} = 0$, hence p_0 is a stationary solution to the forward Kolmogorov equation with differential operator $L - L_1$. In particular this means that $p_t = p_0$ for all positive time.

Now, (17) (and its weak formulation (18)) are precisely the Fokker-Planck equation for the *residual* SDE, which is the linear additive noise SDE given by

$$dX_t = \bar{A}X_t dt + \bar{G}dW_t, \tag{19}$$

such that $\bar{G}\bar{G}^T = \bar{H}$. We have already seen that solutions to (18) are necessarily stationary; equivalently, $X_t \sim X_0 \ \forall t \geq 0$. Then, we note that if $\bar{H} \neq 0$, (19) can only admit a Gaussian stationary distribution. In particular, stationarity would require that (19) is an Ornstein-Uhlenbeck process [Doo42], whose stationary distribution is $\mathcal{N}(0, \Sigma)$, where $\Sigma\bar{A} + \bar{A}\Sigma = -\bar{H}$ [MHB16]. Thus, if X_0 is not auto-rotationally invariant, then X_0 is non-Gaussian, and we must conclude that $\bar{H} = H - H_1 = 0$. Furthermore, this implies that the residual SDE is deterministic, such that

$$\frac{dX_t}{dt} = \bar{A}X_t \implies X_t = e^{\bar{A}t}X_0 \tag{20}$$

$$X_t = e^{\bar{A}t}X_0 \sim X_0 \quad \forall t \geq 0. \tag{21}$$

However, note that $e^{\bar{A}t}X_0 \sim X_0$ admits only the trivial solution $\bar{A} = 0$ since X_0 is not auto-rotationally invariant. This proves that $A = A_1$ and $H = H_1$ as desired. □

Theorem 1 tells us that identifiability from marginals is guaranteed if the initial distribution X_0 is not auto-rotationally invariant. Conversely, if X_0 is auto-rotationally invariant, then it is possible to construct multiple SDEs with the same ensuing marginals. However, we note that in practice, observed marginals may not be compatible with our constructed example (11). An open question is if non-identifiability persists given any set of observed marginals, which start from an auto-rotationally invariant distribution X_0.

As discussed previously, most distributions are not auto-rotationally invariant, and are hence conducive to identifiability. Figures 2a, 2b, and 2c demonstrate how the congruence of temporal marginals from the unidentifiable SDE pairs in Section 4 is broken by various initial distributions X_0.

4.2 Causal graph identification

Given that the drift and diffusion are identified, we can gain important insights into the system's causal dynamics. It has already been shown that knowing drift and diffusion provides the system's post-intervention distributions [HS14, WGH+24]. We prove below that under basic conditions, we can recover the causal graph $\mathcal{G}$ of an additive noise SDE from the drift $b(X_t)$ and observational diffusion $H = GG^T$. In particular, we will see that the simple directed edges $e = i \to j$ in $\mathcal{G}$, indicating a causal effect of $X^{(i)}$ on $X^{(j)}$, are determined by the drift $b(X_t) = \begin{bmatrix} b_1(X_t) & \dots & b_d(X_t) \end{bmatrix}^T$. Similarly, the multidirected edges $\tilde{e} = (i_1, \dots i_p)$, indicating a latent confounder causing variables $X^{(i_1)}, \dots, X^{(i_p)}$ are determined by the diffusion G.

Proposition 1. *Let X_t evolve according to a d-dimensional additive noise SDE: $dX_t = b(X_t)dt + GdW_t$ (2), and $\mathcal{G} = (V = [d], E, \tilde{E})$ be its causal graph, over the variables $X^{(1)}, \cdots, X^{(d)}$, with directed edge set E and multidirected edge set $\tilde{E}$.*

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

b. There exists a multidirected edge $\tilde{e}$ containing (i, j) in $\mathcal{G}$, corresponding to a latent confounder of $X^{(i)}$ and $X^{(j)}$, if and only if $H_{i,j} = (GG^T)_{i,j} \neq 0$.

Thus, if there are no unobserved confounders causing more than two observed variables, then $\mathcal{G}$ is fully identified by $b(X_t)$ and $H = GG^T$.

Proof. Since the SDE has additive noise, the dynamic structural causal model [BM24] is given by

$$X_t^{(j)} - X_0^{(j)} = \int_0^t b_j(X_s)ds + \sum_{k=1}^m G_{j,k} \int_0^t dW_s^{(k)} \tag{22}$$

$$= \int_0^t b_j(X_s)ds + \sum_{k=1}^m G_{j,k}\tilde{\epsilon}_t^{(k)} \tag{23}$$

$$:= f(pa(X_t^{(j)}) + \epsilon_t^{(j)}, \tag{24}$$

where $\tilde{\epsilon}_t^{(k)} \sim \mathcal{N}(0, t)$ are independent noise samples, and $\epsilon_t^{(j)} \perp\!\!\!\perp \epsilon_t^{(i)}$ if and only if $G_i \cdot G_j = 0$. This is due to multivariate Brownian motion being jointly Gaussian, and the property that jointly Gaussian distributions are independent if and only if they are uncorrelated.

From Definition 1 in [BM24], we have the directed edge $i \to j$ in $\mathcal{G}$ if the right hand side of (24) depends on $X^{(i)}$, either via the drift $b_j(X_s)$ or the diffusion. Since additive noise diffusion is independent of X_t, then $i \to j$ if and only if $b_j(X_t)$ depends on $X_t^{(i)}$. This proves claim *a*.

Then, as defined in [LRW21], we include the multidirected edge $(i_1, \dots i_p)$ in $\mathcal{G}$ if variables $X^{(i_1)}, \dots, X^{(i_p)}$ share a latent confounder. This occurs precisely when $X^{(i_1)}, \dots, X^{(i_p)}$ have correlated noise from a common noise source [MH22] (see also [BM24][Example 4] or [HSKP08]). In particular, for an additive noise model, this corresponds to the existence of a column $G_{\cdot,k}$ in the diffusion, such that $G_{i_1,k}, \dots G_{i_p,k} \neq 0$. In the case of two variables, $X^{(i)}$ and $X^{(j)}$, sharing a noise source, this condition is equivalent to $H_{i,j} = G_i \cdot G_j \neq 0$, since there would be some $k \in [m]$ such that $G_{i,k}, G_{j,k} \neq 0$. Thus, $X^{(i)}$ and $X^{(j)}$ would share noise $\epsilon_t^{(k)}$ from the source $W_t^{(k)}$. We conclude that there is a multidirected edge containing (i, j) in $\mathcal{G}$ iff $H_{i,j} \neq 0$. This proves claim *b*.

□

Although the observational diffusion $H = GG^T$ informs the presence of latent confounders between pairs of variables, we note that it is impossible for H alone to determine general multidirected edges $(i_1, \cdots, i_p)$ over $p > 2$ components, which indicate an unobserved confounder of $X_t^{(i_1)}, \cdots, X_t^{(i_p)}$. Indeed, given only $H = GG^T$, multiple causal interpretations may be possible, as shown in Appendix Example 1, where we cannot distinguish between a single multidirected edge $(1, 2, 3)$, and three pairs of bidirected edges $(1, 2), (1, 3), (2, 3)$.

Despite this difficulty, Proposition 1 highlights the power of drift-diffusion identification towards recovering the causal graph. The setting accommodates any stochastic process with additive noise, and is thus equipped to handle complex deterministic relationships between variables, e.g. the interactions between genes in GRNs are commonly modeled via additive noise SDEs [WWT+18, WTW+23, RCM+24, Zha24, LZKS21, CZHS22]. Knowing the drift and diffusion of an additive noise process allows us to identify all simple directed edges in the graph, including cycles, which model feedback loops in GRNs. Even in the worst case, given drift and diffusion, we can recover the graph up to the distinction between multidirected edges and sets of bidirected edges. For an in depth overview of dynamic structural causal models and their relation to defining causal graphs, we refer the reader to Appendix A.1 and references [HS14, MH22, BM24].

We conclude this section with a few remarks related to conditions for learning the system's full causal structure. If we assume that the data obeys causal sufficiency (no latent confounders), and follows a linear additive noise model, such that the initial distribution is not auto-rotationally invariant, then combining Theorem 1 and Proposition 1 shows that we can identify the full causal structure. This is analogous to the celebrated causal identifiability result from [SHH+06] in the static setting. They showed that the full causal structure can be identified if the data obeys causal sufficiency, follows a linear model, and has non-Gaussian

additive noise. We note that Gaussians are a subset of auto-rotationally invariant distributions. The reason why we have a slightly larger class of pathological distributions in the dynamic case is due to the inability to freely scale the mean and covariance of the data, as done in the static setting, to obtain independent components via independent component analysis (ICA). Indeed, scalings of different proportions would be required at each time step, in order to preserve the SDE, as evidenced by Ito's Lemma. For this reason, we cannot apply ICA in the same manner as in the static case, and therefore the conditions for Maxwell's Theorem, which imply Gaussianity, do not hold.

5 Our parameter estimation method

In this section, we introduce Alternating Projection Parameter Estimation from X_0 (APPEX) which can infer drift, diffusion, and trajectories from a set of observed temporal marginals, without prior knowledge. In this work, we focus on the case where the noise is additive and the drift $b(X_t) = AX_t : A \in \mathbb{R}^{d\times d}$, is linear (2), but the algorithm is applicable to any parametric family of drifts, provided that a maximum likelihood estimator (MLE) is used.

Given temporal marginals $p_0, ..., p_{N-1}$ and an initial guess for drift and diffusion $(A^{(0)}, H^{(0)})$, the idea is to use an alternating optimization algorithm to obtain increasingly better estimates for the drift and diffusion matrices. We consider the spaces

$$\mathcal{D} = \{q : q_0 = p_0, ..., q_{N-1} = p_{N-1}\} \tag{25}$$

$$\mathcal{A} = \{p : \exists (A, H) \in \mathbb{R}^{d\times d} \times Sym_d \text{ s.t. } p(x, t|y, s) \sim \mathcal{N}(e^{A(t-s)}y, H(t-s))\}, \tag{26}$$

where $\mathcal{D}$ is the set of laws on paths sharing the N temporal marginals and $\mathcal{A}$ is the set of laws on paths given by linear additive noise SDEs. In practice, the temporal marginals $p_0, \dots, p_{N-1}$ will be empirical distributions for the marginals of the true SDE law at N given times; in the limit of infinite data per time point, the temporal marginals $p_0, \dots, p_{N-1}$ will be the true marginal distributions of the law of the SDE being observed. Each iteration of our method will alternate between information projections, $\arg\min_{q\in\mathcal{D}} D_{KL}(q\|p)$, which represent trajectory inference; and moment projections, $\arg\min_{p\in\mathcal{A}} D_{KL}(q\|p)$, which correspond to the infinite-data limit of maximum likelihood parameter estimation [Mad12]. We note that KL divergence D_{KL} is also commonly known as *relative entropy*. Assuming perfect implementation, an iteration entails the updates:

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

$$A^{(k)} = \arg\min_{A\in\mathbb{R}^{d\times d}} D_{KL}(q^{(k)}\|p_{A,H^{(k-1)}}) \tag{28}$$

$$H^{(k)} = \arg\min_{H\in Sym_d} D_{KL}(q^{(k)}\|p_{A^{(k)},H}) \tag{29}$$

At iteration k, we first use the previous iteration's estimated SDE parameters to define a reference SDE $p_{A^{(k-1)},H^{(k-1)}}$. We then perform trajectory inference (27), by determining the law on paths $q^{(k)} \in \mathcal{D}$, which minimizes relative entropy to the reference SDE, while also satisfying the marginal constraints (25). This step amounts to solving a version of the Schrodinger Bridge problem [VTLL21, BHCK23]. After an appropriate law on paths $q^{(k)}$ is obtained, we project $q^{(k)}$ back onto $\mathcal{A}$, to find the law of a linear additive noise SDE $p_{A^{(k)},H^{(k)}}$ that minimizes relative entropy to $q^{(k)}$. This moment projection is performed in two steps via maximum likelihood estimation with respect to $q^{(k)}$. We first update the estimated linear drift in (28) by finding the MLE $A^{(k)} \in \mathbb{R}^d$, given that the diffusion is $H^{(k-1)}$. Finally, we update the diffusion estimate in (29), by finding the MLE diffusion $H^{(k)} \in Sym_d$, given that the drift matrix is $A^{(k)}$. We note that an analogous procedure can be used for a general additive noise SDE (1), provided that a corresponding maximum likelihood estimator is used in the drift update step for the estimated nonlinear drift function $b_k(X_t)$ (28). Indeed, we show in Corollary 2 that the structure of the closed form diffusion MLE is unchanged for general additive noise SDEs.

Previous works [Zha24, SBB24, VTLL21] implement similar iterative schemes, which alternate between trajectory inference (Schrodinger Bridge problem) and maximum likelihood 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 theoretical reason for this assumption is to ensure finite KL divergence [VTLL21].

Given continuously observed marginals of a d-dimensional process, the KL divergence between two laws on paths q, k is taken over the path space $\Omega = C([0,T], \mathbb{R}^d)$, such that

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

Thus, $D_{KL}(q\|p)$ will only be finite between two laws on paths if we can define the Radon-Nikodym derivative $\frac{dq}{dp}$ over Ω. By Girsanov's theorem, this is only ensured when both processes share the same diffusion [SBB24, VTLL21]. However, if we only consider measurements from a finite number of observed marginals, the KL divergence of the discretized processes

$$D_{KL}(q\|p) = \sum_{i=0}^{N-1} D_{KL}(q_{t_i,t_{i+1}}\|p_{t_i,t_{i+1}})$$

will be finite as long as for each i, $D_{KL}(q_{t_i,t_{i+1}}\|p_{t_i,t_{i+1}}) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \log\left(\frac{dq_{t_i,t_{i+1}}}{dp_{t_i,t_{i+1}}}(x)\right) q_{t_i,t_{i+1}}(x)dx < \infty$. This is ensured as long as for each time t_i, the joint distributions $q_{t_i,t_{i+1}}$ and $p_{t_i,t_{i+1}}$ are absolutely continuous with respect to one another. In particular, if p and q are the laws of two different drift-diffusion SDES with different diffusion matrices, if the two different diffusion matrices are both non-degenerate then $D_{KL}(q\|p) < \infty$ when we consider the discretized laws on paths over the N observed time points. 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 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 generalized entropic optimal transport

Given two marginals μ, ν, the standard entropy regularized optimal transport (EOT) problem of transporting probability measure a to b with entropic regularization $\epsilon > 0$ and cost $c(x,y) = \|y-x\|^2/2$ is solved via

$$\pi^* = \inf_{\pi\in\Pi(\mu,\nu)} \int c(x,y)d\pi(x,y) + \epsilon^2 D_{KL}(\pi\|\mu\otimes\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int \frac{\|y-x\|^2}{2\epsilon^2} d\pi(x,y) + D_{KL}(\pi\|\mu\otimes\nu). \quad (30)$$

This problem also admits a dual formulation, in terms of finding a pair of potentials (f,g) with respect to the Gaussian transition kernel $K(x,y) \propto e^{\frac{c(x,y)}{\epsilon^2}} \propto e^{\frac{-\|y-x\|^2}{2\epsilon^2}}$ [Jan21, Nut21]:

$$\pi^*(x,y) = e^{f^*(x)+g^*(y)} K(x,y)d\mu(x)d\nu(y) \quad (31)$$

$$f^*, g^* = \sup_{f\in\mathcal{L}^1(\mu), g\in\mathcal{L}^1(\nu)} \mathbb{E}_\mu(f) + \mathbb{E}_\nu(g) - \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{f(x)+g(y)} K(x,y)d\mu(x)d\nu(y) - 1\right). \quad (32)$$

In particular, given marginals μ, ν and the transition kernel K, Sinkhorn's algorithm uses the dual formulation (32) to find the ϵ-entropy regularized optimal transport solution π^*, by solving for f^*, g^* via iterative projections [PC19]. Furthermore, it is easy to see from (32) that solving for the entropic optimal transport solution can be translated to a Schrodinger bridge problem (27), in the sense that $\pi^* \in \Pi(\mu,\nu)$ is the joint density with marginals μ, ν, which minimizes relative entropy with respect to the reference measure K [PC19, Zha24]:

$$\pi^* = \text{Proj}^{KL}_{\Pi(\mu,\nu)}(K) = \operatorname*{arg\,min}_{\pi\in\Pi(\mu,\nu)} D_{KL}(\pi\|K). \quad (33)$$

In the standard ϵ-regularized OT problem (30), $K \sim \mathcal{N}(x, \epsilon^2)$ is an isotropic Gaussian kernel with standard deviation ϵ. As noted in [LZKS21], 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 $P(X_t = \mu, X_{t+dt} = \nu)$, which minimizes relative entropy to the law of a pure diffusion process $dX_t = \sigma dW_t$, one should set the entropic regularization such that $\epsilon^2 = \sigma^2 dt$. Indeed, this would correspond to minimizing the KL divergence to $K(x,y) = e^{-\frac{\|y-x\|^2}{2\sigma^2 dt}} \sim \mathcal{N}(x, \sigma^2 dt)$, which is the transition kernel of the reference SDE. Similarly, as done in [Zha24], one can perform trajectory inference given an Ornstein-Uhlenbeck reference SDE $dX_t = -AX_t dt + \sigma dW_t$ by approximating the transition kernel via $K(x,y) = e^{-\frac{\|y-e^{Adt}x\|^2}{2\sigma^2 dt}}$. This would correspond to reweighting the squared Euclidean cost with the drift

matrix A, such that $c(x,y) = \|y - e^{Adt}x\|/2$, and applying standard entropy regularized OT with $\epsilon^2 = \sigma^2 dt$. Indeed, the problem can be approximated via (30) by adjusting the first marginal $\mu \to e^{Adt}\mu$, and considering a pure diffusion reference SDE $d\tilde{X}_t = \sigma dW_t$ controlling the evolution of $\tilde{X}_0 = e^{Adt}\mu$ to $\tilde{X}_{dt} = \nu$.

However, the standard entropy regularized OT problem (30) only considers a scalar regularization parameter ϵ, and thus can only accommodate reference SDEs whose diffusion is proportional to Brownian motion. To generalize trajectory inference for SDEs with non-isotropic diffusion, we formalize "generalized entropic optimal transport", by parameterizing the transition kernel K in the dual problem with A and $H = GG^T$ rather than the scalar $\epsilon^2 > 0$. Indeed, we can consider generalized Gaussian transition kernels with custom means and covariance, $K_{\theta,\Sigma}(x,y) = exp(\frac{(y-f(x,\theta))^T\Sigma^{-1}(y-f(x,\theta))}{2})$, in order to model a transition $y|x \sim \mathcal{N}(f(x,\theta),\Sigma)$. Under this formulation, the cost is given by the inner product $c(x,y,A,\Sigma) = \frac{(y-f(x,\theta))^T\Sigma^{-1}(y-f(x,\theta))}{2}$, and we solve the generalized entropic OT problem:

$$\pi^* = \inf_{\pi\in\Pi(\mu,\nu)} \int \frac{(y-f(x,\theta))^T\Sigma^{-1}(y-f(x,\theta))}{2} d\pi(x,y) + D_{KL}(\pi\|\mu\otimes\nu), \tag{34}$$

where the entropic regularization is captured in the possibly non-isotropic covariance Σ. Given marginals μ,ν, and the kernel $K_{\theta,\Sigma}$, Sinkhorn's algorithm would find the coupling $\pi^* \in \Pi(\mu,\nu)$, which minimizes relative entropy to $K \sim \mathcal{N}(f_\theta(x),\Sigma)$.

In order to state the next result, we need the notion of *Markov concatenation* of couplings. To wit, given Polish spaces X_1, X_2, X_3 and couplings $\pi_{12} \in \mathcal{P}(X_1 \times X_2)$ and $\pi_{23} \in \mathcal{P}(X_2 \times X_3)$ with identical marginals μ_2 on X_2, the *Markov concatenation* $\pi_{12} \circ \pi_{23}$ of π_{12} and π_{13} is a multi-coupling in $\mathcal{P}(X_1 \times X_2 \times X_3)$ given by

$$\pi_{12} \circ \pi_{23}(dx_1, dx_2, dx_3) = \pi_{12}(dx_1 \mid x_2)\mu_2(dx_2)\pi_{23}(dx_3 \mid x_2).$$

Here $\pi_{12}(dx_1 \mid x_2)$ is the disintegration of π_{12} with respect to μ_2 and $\pi_{23}(dx_3 \mid x_2)$ is the disintegration of π_{23} with respect to μ_2. The interpretation of the Markov coupling is as follows: a random "trajectory" according $\pi_{12} \circ \pi_{23}$ 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 Schrodinger bridges [LZKS21]. 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 $\pi_{12} \circ \pi_{23} \circ \pi_{34}$, see [BCDMN19] Section 3.2.

Proposition 2 (Optimality of Generalized entropic optimal transport)**.** *Let the reference SDE be a linear additive noise SDE with drift A and diffusion H (2). Given a set of observed marginals $p_0, \ldots, p_{N-1}$ over times $\{t_i\}_{i=0}^{N-1}$, let π denote the concatenated joint distribution given by*

$$\pi = \pi_0 \circ \ldots \circ \pi_{N-2},$$

where π_i is the generalized entropic OT solution obtained by Sinkhorn's algorithm with marginals $\mu = p_i, \nu = p_{i+1}$ and transition kernel

$$K^i_{A,H}(x,y) \propto exp(-\frac{1}{2}(y - e^{Adt}x)^T(\Sigma_i)^{-1}(y - e^{Adt}x) \tag{35}$$

where $\Sigma_i = \int_{t_i}^{t_{i+1}} e^{A(t_{i+1}-s)} H e^{A^T(t_{i+1}-s)}\, ds$. Then, π minimizes relative entropy to the law of the reference SDE:

$$\pi = \operatorname*{arg\,min}_{\pi\in\Pi(p_0,\ldots,p_{N-1})} D_{KL}(\pi\|p_{A,H}),$$

where $p_{A,H}$ is the law of the reference SDE, discretized over $p_0, \ldots, p_{N-1}$.

Proof. For each $i = 0, \ldots, N-1$, by (33), we have $\pi_i = \arg\min_{\pi\in\Pi(p_i,p_{i+1})} D_{KL}(\pi\|K^i_{A,H})$. Since $K^i_{A,H} \sim \mathcal{N}(e^{Adt}X_t, \Sigma_i)$ is the transition kernel of the SDE at time t_i, this implies that $\pi_i = \arg\min_{\pi\in\Pi(p_i,p_{i+1})} D_{KL}(\pi\|p_{A,H})$, where the KL divergence is taken over two time points t_i, t_{i+1}. The details are as follows. Let $K_{t_{i+1}-t_i}(x,y)$ denote the transition kernel from time t_i to time t_{i+1} for the SDE $dX_t = AX_t + GdW_t$. In particular, if p_i is the marginal of this SDE at time t_i, then we have that $K_{t_{i+1}-t_i}(x,y)p_i(dx)dy$ is equal to the joint distribution γ_i of p_i and p_{i+1}. Now, the "chain rule" for the relative entropy tells us that if $\gamma, \pi \in \mathcal{P}(X_1 \times X_2)$,

and we write γ_1 and π_1 for their projections onto the first coordinate, and $\gamma(\cdot \mid x_1)$ and $\pi(\cdot \mid x_1)$ for their disintegrations with respect to the first coordinate, we have

$$\mathrm{KL}(\pi \mid \gamma) = \mathrm{KL}(\pi_1 \mid \gamma_1) + \int_{X_1} \mathrm{KL}(\pi(\cdot \mid x_1) \mid \gamma(\cdot \mid x_1)) d\pi_1(x_1).$$

Accordingly, we have that when $\gamma = K_{t_{i+1}-t_i}(x,y)p_i(dx)dy$,

$$\min_{\pi \in \Pi(p_i, p_{i+1})} \mathrm{KL}(\pi \mid \gamma) = \mathrm{KL}(p_i \mid p_i) + \min_{\pi \in \Pi(p_i, p_{i+1})} \int_{\mathbb{R}^d} \mathrm{KL}(\pi(\cdot \mid x) \mid K_{t_{i+1}-t_i}(x,y)dy) dp_i(x)$$

while at the same time, if we instead take $\gamma' = K_{t_{i+1}-t_i}(x,y)dxdy$, we get

$$\min_{\pi \in \Pi(p_i, p_{i+1})} \mathrm{KL}(\pi \mid \gamma') = \mathrm{KL}(p_i \mid dx) + \min_{\pi \in \Pi(p_i, p_{i+1})} \int_{\mathbb{R}^d} \mathrm{KL}(\pi(\cdot \mid x) \mid K_{t_{i+1}-t_i}(x,y)dy) dp_i(x).$$

Thus the minimizers for the two minimization problems are identical.

Since π is constructed as a Markov concatenation, the conclusion follows from Lemma 3.4 of [BCDMN19], 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 contatenation $\pi_1 \circ \ldots \pi_{N-2}$, and has ith marginal equal to p_i, we have

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

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

□

We note that $X_{t+dt}|X_t \sim \mathcal{N}(e^{Adt}X_t, \Sigma_t)$ can practically be estimated via the first order approximation $\mathcal{N}(X_t + AX_t dt, Hdt)$.

5.2 Parameter estimation via MLE

To optimize objectives (28) and (29) for each iteration of APPEX, we require maximum likelihood estimators for the SDE parameters in the setting of multiple observed trajectories from $[0, T]$. We derive closed-form maximum likelihood solutions for the linear additive noise SDE from Equation 2, given multiple trajectories. 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.

Proposition 3 (MLE estimators for drift and diffusion of SDE (2) from multiple trajectories)**.** *Given M trajectories over N different times: $\{X^{(j)}_{t_i} : i \in 0, ..., N-1,\ j \in 0, ..., M-1\}$ sampled from the linear additive noise SDE (2), the maximum likelihood solution for linear drift is approximated by*

$$\hat{A} = \frac{1}{dt} \left(\sum_{i=0}^{N-2} \sum_{j=0}^{M-1} \Delta X^{(j)}_i X^{(j)^T}_i \right) \left(\sum_{i=0}^{N-2} \sum_{j=0}^{M-1} X^{(j)}_i X^{(j)^T}_i \right)^{-1} \tag{36}$$

and the maximum likelihood solution for diffusion is approximated by

$$\hat{H} = \frac{1}{MT} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} \left(\Delta X^{(j)}_i - AX^{(j)}_i dt)(\Delta X^{(j)}_i - AX^{(j)}_i dt)^T \right) \tag{37}$$

Proof. See Section (A.5) of the Appendix. We note that estimators $\hat{A}$ and $\hat{H}$ were derived using the discretized transition kernel, $X_{i+1}|X_i \sim \mathcal{N}(X_i + AX_i dt, Hdt)$. We derive the maximum likelihood estimators with the exact transition kernel $X_{i+1}|X_i \sim \mathcal{N}(e^{Adt}X_i, Hdt)$ in the one dimensional case in Section (A.5). □

Remark 1. *Note that the MLE estimator for drift A does not depend on the diffusion H, but the MLE estimator for H depends on A. Therefore, we estimate drift first in each iteration of APPEX.*

5.3 The APPEX algorithm

To reiterate, our novel algorithm APPEX has three primary subprocedures. In the first step, we perform a trajectory inference step as outlined in Equation 27 (step 6 in Algorithm 1). Second, we find the MLE for drift as outlined in Equation 28 (step 9 in Algorithm 1). Third, we update our diffusion estimate via Equation 29 (step 10 in Algorithm 1). Given the optimality of each of APPEX's subprocedures, as shown in the previous subsections, it follows that in the limit where the marginals $p_0, \dots, p_{N-1}$ are observed exactly, the drift and diffusion estimates are improving with each iteration, and getting closer to the coefficients underlying SDE [SBB24]. Indeed, by construction, we have

$$D_{KL}(q^{(k+1)} \| p_{A^{k+1},H^{k+1}}) \leq D_{KL}(q^{(k+1)} \| p_{A^{k+1},H^k}) \leq D_{KL}(q^{(k+1)} \| p_{A^k,H^k}) \leq D_{KL}(q^{(k)} \| p_{A^k,H^k})$$

Let the true parameters be given by A, H. Since $p_{A,H} \in \mathcal{D}$, then $\inf_{q \in \mathcal{D}, p \in \mathcal{A}} D_{KL}(q\|p) = 0$. Furthermore, if X_0 is not auto-rotationally invariant, then $\mathcal{A} \cap \mathcal{D} = \{p_{A,H}\}$, and hence

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

In this sense, our method is always approaching the true solution. However, due to the non-convexity of the set $\mathcal{A}$, classical optimization theory does not apply, and we cannot use these results to prove convergence to the true solution.

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

Input: Observed marginals p_i, $i = 0, \dots, N-1$, number of iterations K, Δt
Result: Estimated drift function $\hat{b}$ and additive noise $\hat{H}$
$\hat{b} \leftarrow 0$, $H \leftarrow I$, $k \leftarrow 0$
while $k < K$ **do**
  for $i = 1, \dots, N$ **do**
    $\pi_{i,i+1} \leftarrow$ Generalized-Entropic-Optimal-Transport$(\hat{b}, \hat{H}, p_{i-1}, p_i, \Delta t)$
  end for
  Sample-Trajectories $\leftarrow \pi_{N-1,N} \circ \dots \circ \pi_{1,2}(p_0)$
  $\hat{b} \leftarrow$ MLEfit(Sample-Trajectories)
  $\hat{H} \leftarrow$ MLEfit(Sample-Trajectories, $\hat{b}$)
  $k \leftarrow k + 1$
end while
$\mathcal{G} \leftarrow$ Estimate-Causal-Graph$(\hat{b}, \hat{H}, \epsilon)$

As proven in Proposition 1 and visualized in Figure 1, the causal graph can be derived from the SDE parameters. For nonlinear drift $b(X_t)$, one can perform independence testing to check $b_j(X_t) \not\perp\!\!\!\perp X_t^{(i)}$ according to some threshold $\epsilon > 0$, in order to determine the edge $i \to j$. For linear drift one may simply examine $A_{j,i}$. For multi-edges, one can check $H_{i,j}$ to determine whether there is an unobserved confounder causing $X^{(i)}$ and $X^{(j)}$, represented by a multidirected edge containing (i, j). We acknowledge that many causal algorithms and independence testing are applicable, and leave the implementation general.

6 Experiments

In this section, we first demonstrate that the non-identifiable examples from Section (3) can be resolved if X_0 is not auto-rotationally invariant. To show this, we demonstrate that our parameter estimation method APPEX from Section 5 effectively estimates the drift and diffusion parameters following an appropriate non-auto-rotationally invariant initialization. We then show that combining our initialization with our parameter estimation method successfully learns arbitrary dynamical systems modelled with a linear additive noise SDE by testing it on randomly generated SDEs across a range of dimensions. We conclude this section with several causal graph recovery experiments.

6.1 Experimental setup

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 $p_0 \sim unif\{x_i\}_{i=1}^d$ where $\{x_i\}_{i=1}^d$ are randomly sampled linearly independent vectors with each entry having a magnitude between 2 and 10. As proven by Proposition 4 and Theorem 1, this initial distribution ensures that parameter estimation is feasible.

To model the setting where we only observe temporal marginals, we subsample at the rate $dt = 0.05$ to produce $N = 20$ marginals with 500 observations per time. To perform parameter estimation, we use 30 iterations of our APPEX algorithm, such that the initial reference SDE is an isotropic Brownian motion $dX_t = \sigma dW_t$, even when the true process has non-isotropic 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 $Tr(H)$. Specifically, we initialize $A^{(1)} = 0, H^{(1)} = \sigma^2 I$ $s.t.$ $\sigma^2 \sim tr(H)10^{Unif(-1,1)}$. Because the first reference SDE is σdW_t, we note that the first iteration of APPEX is equivalent to inferring trajectories using the Waddington-OT (WOT) method from [SST+19], based on standard regularized OT (30) with $\epsilon^2 = \sigma^2 dt$, followed by MLE parameter estimation. APPEX and WOT are distinguished by the fact that APPEX benefits from further iterations, and allows the reference SDE to have non-isotropic diffusion and non-zero drift. APPEX's subprocedures are implemented using our generalized entropic optimal transport method from Proposition (2) and the MLEs found in Proposition (3). Due to time complexity and numerical stability, we use linearized discretizations for the Gaussian transition kernels except for dimension 1.

6.2 Revisiting previously non-identifiable SDEs

We perform parameter estimation on the three pairs of SDEs from Section 4. 10 replicates of each experiment were performed, such that each replicate featured data from different valid p_0 initializations and different initial diffusivities σ^2. To measure performance, we track the mean absolute error (MAE), plotted in Figure 3, between the true drift/diffusion parameters and their estimates.

The results demonstrate that in each example, APPEX is able to estimate both drift and diffusion by iteratively improving upon both estimates. In each experiment, we observe decreasing MAEs for both SDE parameters as the number of iterations increases. The worst performance was observed for the SDE $dX_t = -10X_t dt + \sqrt{10} dW_t$, whose significantly higher noise scale made inference more difficult, particularly for the diffusion, which APPEX consistently estimated around $H = 7$ rather than $H = 10$.

We empirically demonstrate in Section B.1 that over a broad class of random experiments, APPEX's estimates for both drift and diffusion converge to the true SDE parameters. We further note that even though degenerate diffusions can result in infinite KL divergence with respect to a mispecified reference SDE, we also observe that APPEX can estimate degenerate diffusions, such as in example 9, without prior knowledge.

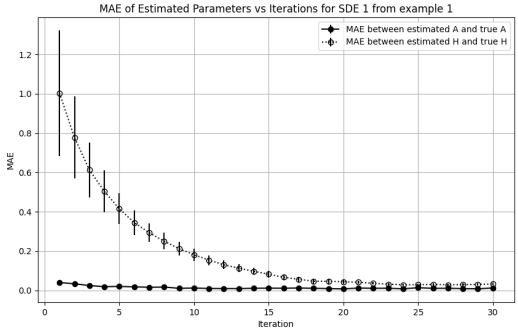
6.3 Higher dimensional random matrices

We now test APPEX 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 initialize each of its d^2 entries via $Unif(-5, 5)$. To ensure that the system does not blow up, we verify that the maximal real part of the eigenvalues of A is less than 1. This allows us to consider process beyond OU processes, while obeying reasonable growth conditions in practice. To create each diffusion G, we randomly initialize each of its d^2 entries via $Unif(-1, 1)$ and then set $H = GG^T$. For numerical stability, we perform Sinkhorn on the logarthmic scale for this experiment.

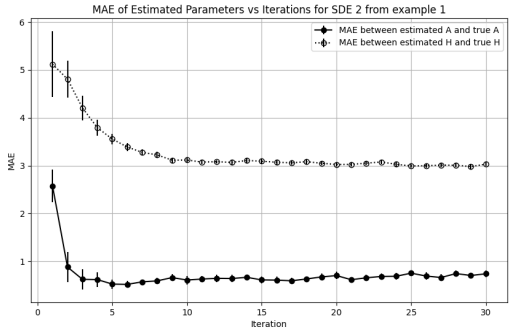
The results demonstrate that APPEX continues to estimate both SDE parameters robustly across all settings. Importantly, APPEX can handle arbitrary additive noise structures $H = GG^T$, as evidenced by high correlations to the true diffusion and low MAE. This is significant because previous literature has focused on the setting of isotropic noise, but in practice, we expect noise structures to have off-diagonal entries due to correlated noise from unobserved confounders, 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 (4) shows how APPEX is able to re-orient incorrect diffusion priors to closely match the true diffusion of the underlying SDE.

6.4 Causal discovery

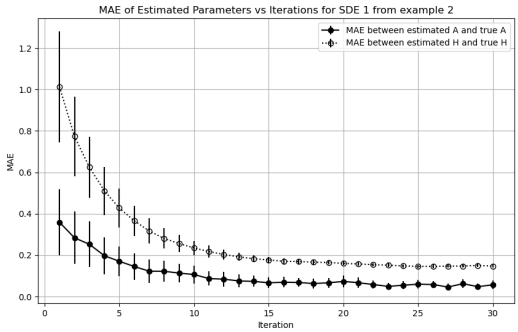
We conclude with experiments that demonstrate APPEX's ability to recover the causal graph of the underlying dynamic system. We first consider causal discovery from systems without latent confounders, and then



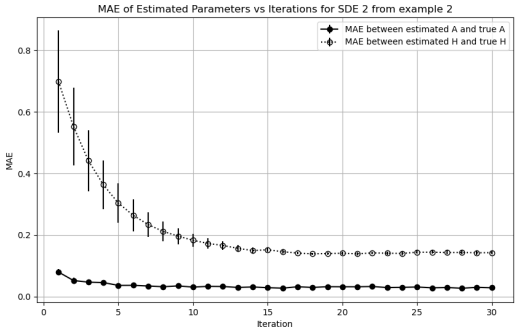
(a) $dX_t = -X_t dt + dW_t$ from Example (5)



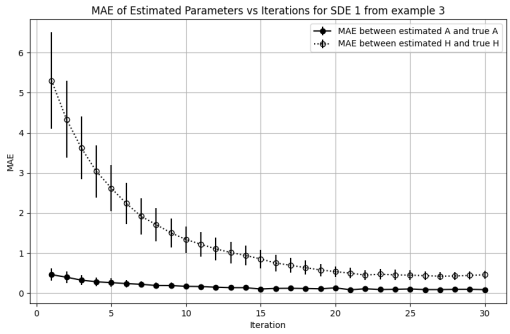
(b) $dX_t = -10X_t dt + \sqrt{10}dW_t$ from (5)



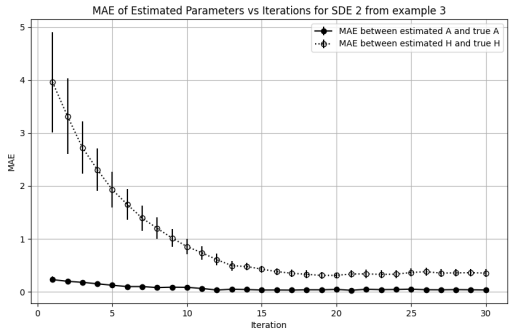
(c) $dX_t = dW_t$ from (7)



(d) $dX_t = \begin{bmatrix} 0 & 1 \\ -1 & -0 \end{bmatrix} X_t dt + dW_t$ from (7)



(e) $dX_t = \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix} X_t dt + \begin{bmatrix} 1 & 2 \\ -1 & -2 \end{bmatrix} dW_t$ from (9)



(f) $dX_t = \begin{bmatrix} \frac{1}{3} & \frac{4}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{bmatrix} X_t dt + \begin{bmatrix} 1 & 2 \\ -1 & -2 \end{bmatrix} dW_t$

Figure 3: The mean absolute error for estimates of A and H using APPEX is shown per iteration for all three pairs of SDEs from Section (4)

consider causal discovery from systems with latent pairwise confounders.

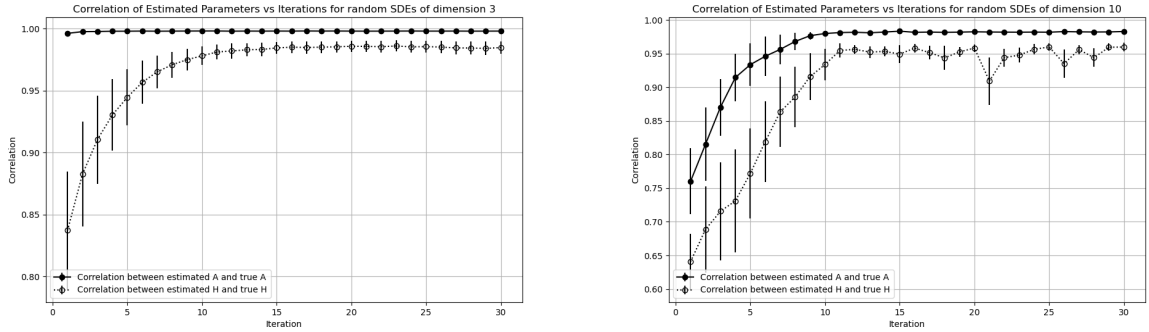
We evaluate APPEX using the default experimental settings laid out in Section 6.1 across randomly generated SDEs of dimension $d = 3, 5, 10$. In particular, we consider Erdös-Renyi graphs $G(d, p)$, such that each of the $d(d-1)$ possible simple directed edges are included with probability p. As in Section 6.3, we ensure that the maximal eigenvalue of the randomly generated ground truth drift matrix is at most 1 to prevent blow-ups.

Table 1: Mean absolute error (MAE) of estimated drift and diffusion for dimensions 3-10. Our method (APPEX) consistently outperforms (bold) the previous method (WOT).

Dimension	A Estimation		GG^T 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 2: Correlation between estimated and true drift and diffusion for dimensions 3-10. Our method (APPEX) consistently outperforms (bold) the previous method (WOT).

Dimension	A Estimation		GG^T 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	$\mathbf{0.960 \pm 0.006}$



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

Figure 4: 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

6.4.1 Causal discovery under causal sufficiency

Proposition 1 states that simple 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 for simulating data for causal discovery, by simulating edge weights $A_{j,i}$ uniformly via $Unif(-5, 0.5) \cup Unif(0.5, 5)$ [Run21, RSW21]. We then determine the presence of 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 $\epsilon = 0.5$ according to the minimal edge weight magnitude from simulated construction.

To model the case where the observed system does not have any latent confounders, we set the diffusion G to be zero for all entries outside its main diagonal. Thus, the only edges in the causal graph $\mathcal{G}$ are simple directed edges, attributed to the drift. Similarly to the experiment on higher dimensional random matrices,

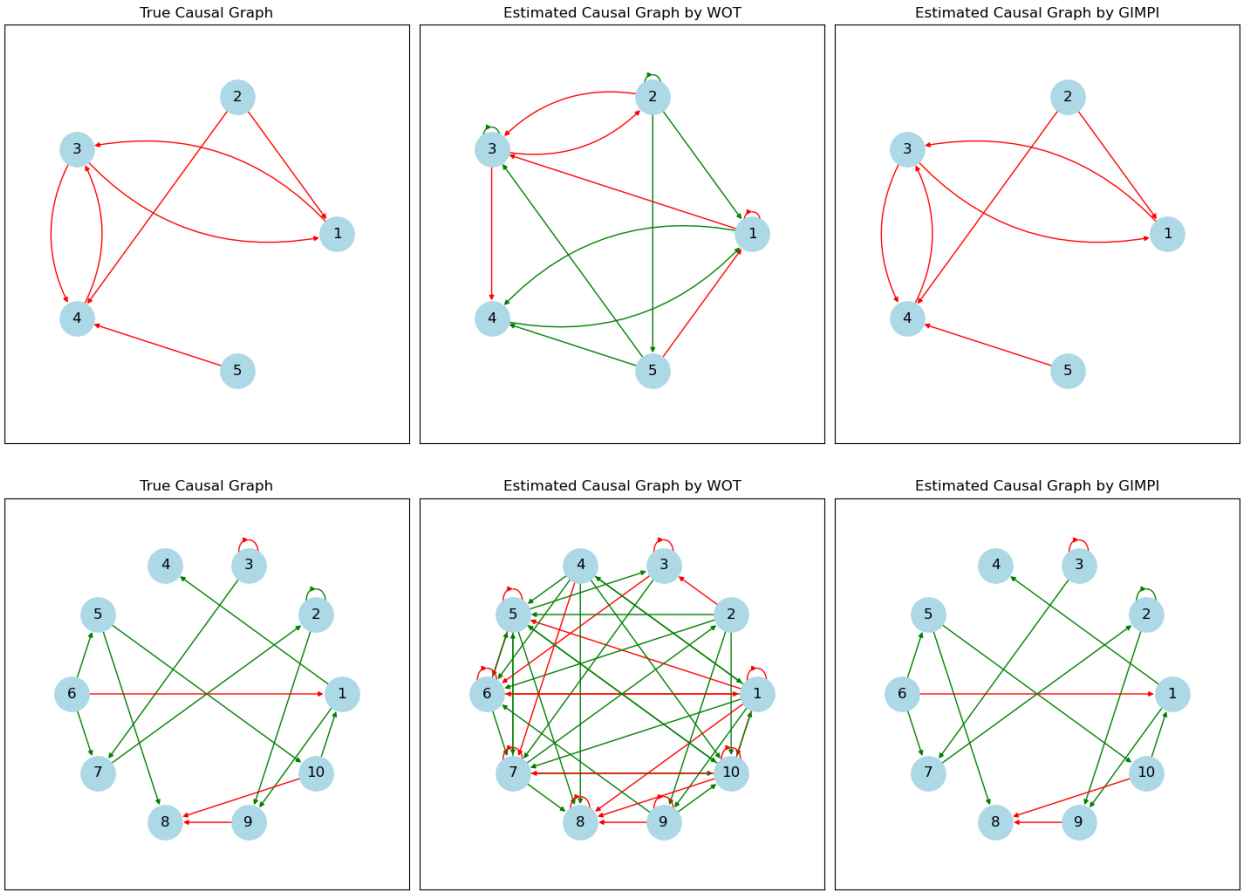


Figure 5: The true and estimated causal graphs by WOT and APPEX for two random SDEs of dimensions $d = 5, 10$ are illustrated. Red edges $i \to j$ represent negative edge weights, such that $X^{(i)}$ negatively regulates $X^{(j)}$. Green edges $i \to j$ represent positive edge weights, such that $X^{(i)}$ positively regulates $X^{(j)}$.

we set each of the d diagonal entries of G via $Unif(0, 1)$.

We measure performance according to the structural Hamming distance. In particular, a system's causal graph consists of directed edges $i \to j$ with positive weights $A_{j,i} > 0$ ($\hat{A}_{j,i} > \epsilon$), directed edges $i \to j$ with negative weights $A_{j,i} < 0$ ($\hat{A}_{j,i} < -\epsilon$), and absence of edges $A_{j,i} = 0$ ($|\hat{A}_{j,i}| < \epsilon$). The structural Hamming distance adds 1 for every instance in which an edge is misclassified, and is defined by

$$d(\mathcal{G}(A), \mathcal{G}(\hat{A})) = \sum_{(i,j)\in[d]\times[d]} \mathbf{1}\left\{\text{sgn}(A_{j,i}) \neq \text{sgn}(\hat{A}_{j,i})\mathbf{1}_{|\hat{A}_{j,i}|>\epsilon}\right\}. \tag{38}$$

The mean structural Hamming distances of causal graphs estimated by WOT and APPEX, across various dimensions $d \in \{3, 5, 10\}$ and random edge probabilities $p \in \{0.1, 0.25, 0.5\}$, are given in Table 3. We also plot the true vs. estimated graphs by WOT and APPEX for two SDEs in Figure 5.

While WOT often struggles to learn the causal graph, due to drift estimates compensating for mispecified diffusion [Hua24], APPEX consistently recovers the causal graph, including cycles and v-structures. For example, the first row in Figure 5 demonstrates that APPEX recovers the system's negative feedback structure, while WOT introduces additional cycles between variables to make sense of the data with poorly estimated diffusion. This highlights the need to estimate diffusion accurately, in order to recover an accurate causal representation of the system from the drift.

6.4.2 Causal discovery with latent confounders

Proposition 1 also states that pairwise latent confounders, represented by bi-directed edges $\tilde{e} = i \leftrightarrow j$ in the causal graph $\mathcal{G} = ([d], E, \tilde{E})$ are characterized by the condition $H_{i,j} = H_{j,i} \neq 0$. Since we would also like to

Table 3: Average Structural Hamming Distance (lower is better) with varying dimensions and random edge probabilities p. Our method (APPEX) consistently outperforms (bold) the previous method (WOT)

Dimension	$p = 0.1$		$p = 0.25$		$p = 0.5$	
	WOT	**APPEX**	**WOT**	**APPEX**	**WOT**	**APPEX**
3	0.40 ± 0.40	$\mathbf{0.00 \pm 0.00}$	0.40 ± 0.16	$\mathbf{0.20 \pm 0.13}$	1.50 ± 0.76	$\mathbf{0.00 \pm 0.00}5$
5	7.30 ± 2.40	$\mathbf{4.10 \pm 1.69}$	6.50 ± 1.88	$\mathbf{0.00 \pm 0.00}$	3.60 ± 1.44	$\mathbf{0.60 \pm 0.22}$
10	47.20 ± 6.28	$\mathbf{0.70 \pm 0.30}$	38.20 ± 8.40	$\mathbf{1.50 \pm 0.43}$	42.5 ± 5.15	$\mathbf{3.70 \pm 0.83}$

consider cycles formed by sets of simple directed edges, we consider the augmented causal $\bar{\mathcal{G}} = (V, E)$ to more clearly model pairwise latent confounders, as done in [BM24]. Specifically, $\tilde{e} = i \leftrightarrow j \in \tilde{E}$ would correspond to the v-structure $i \leftarrow U_{i,j} \rightarrow j$ in the augmented graph $\bar{\mathcal{G}}$, where $U_{i,j}$ is the unobserved pairwise confounder of $X^{(i)}$ and $X^{(j)}$.

To model the case where the observed system features pairwise latent confounders, we randomly select a subset of the columns of the diffusion G to feature precisely two nonzero entries, in two randomly chosen rows i, j. For simplicity, we initialize these entries $G_{i,k} = G_{j,k} = 1 \implies H_{i,j} = G_i \cdot G_j \geq 1$. We then determine the presence of the v-structure $i \leftarrow U_{i,j} \rightarrow j$ if in the estimated observational diffusion, $|\hat{H}_{i,j}| = |\hat{H}_{j,i}| > \epsilon$. We again pick our threshold to be $\epsilon = 0.5$.

We again measure performance according to the structural Hamming distance. For interpretability, we break the distance into two parts. We consider the Hamming distance based on the simple edges between observed variables, defined in (38), as well as the Hamming distance based on latent v-structures from pairwise latent confounders, which adds 1 for every misclassified v-structure in the augmented graph:

$$d(\bar{\mathcal{G}}(H), \bar{\mathcal{G}}(\hat{H}) = \sum_{(i,j) \in [d] \times [d] : i \neq j} \mathbf{1}\left\{ (|H_{i,j}| > \epsilon \cap |\hat{H}_{i,j}| \leq \epsilon) \cup (|H_{i,j}| \leq \epsilon \cap |\hat{H}_{i,j}| > \epsilon) \right\}. \tag{39}$$

We evaluate APPEX and WOT across random SDEs of dimensions $d = 3, 5, 10$ with random edge probability $p = 0.25$ for simple edges. The number of pairwise latent confounders is chosen uniformly from $\{1, \ldots, \lfloor \frac{2d}{3} \rfloor\}$. The results are summarized in Table 4. We also plot the true vs. estimated graphs by WOT and APPEX for two SDEs in Figure 6. We see that APPEX is able to recover both the simple edges and the latent v-structures with high accuracy across dimensions. In constrast, WOT struggles to detect both, especially for higher dimensions. We note that in this setting with non-isotropic diffusion, WOT's initial diffusion reference is not only mispecified according to diagonal scaling, but further mispecifies the nonzero entries.

Table 4: Average Structural Hamming Distance for simple edges, and latent v-structures (lower is better) with varying dimensions and random edge probability $p = 0.25$. Our method (APPEX) consistently outperforms (bold) the previous method (WOT).

Dimension	Hamming distance for simple edges		Hamming distance for latent v-structures	
	WOT	**APPEX**	**WOT**	**APPEX**
3	1.40 ± 0.31	$\mathbf{0.5 \pm 0.27}$	1.70 ± 0.30	$\mathbf{0.20 \pm 0.13}$
5	7.30 ± 1.13	$\mathbf{1.90 \pm 0.67}$	2.90 ± 0.81	$\mathbf{0.20 \pm 0.13}$
10	38.80 ± 5.98	$\mathbf{2.70 \pm 0.67}$	23.10 ± 3.05	$\mathbf{0.40 \pm 0.40}$

7 Conclusions

This is the first work that tackles the estimation of both drift and diffusion of SDEs from marginal snapshots. To do so, we first provide necessary and sufficient conditions for the identifiability of arbitrary linear additive noise SDEs from temporal marginals, which is guaranteed if and only if the initial distribution is not auto-rotationally invariant. We also introduce the first method to fully identify a dynamical system from only marginal observations. To estimate arbitrary drift and additive noise, we generalize entropic optimal transport for non-isotropic diffusions, and we pair this with maximum likelihood estimates of drift and diffusion. Our parameter estimation method, APPEX, then iterates towards the optimal parameter set by alternating between trajectory inference and parameter estimation.

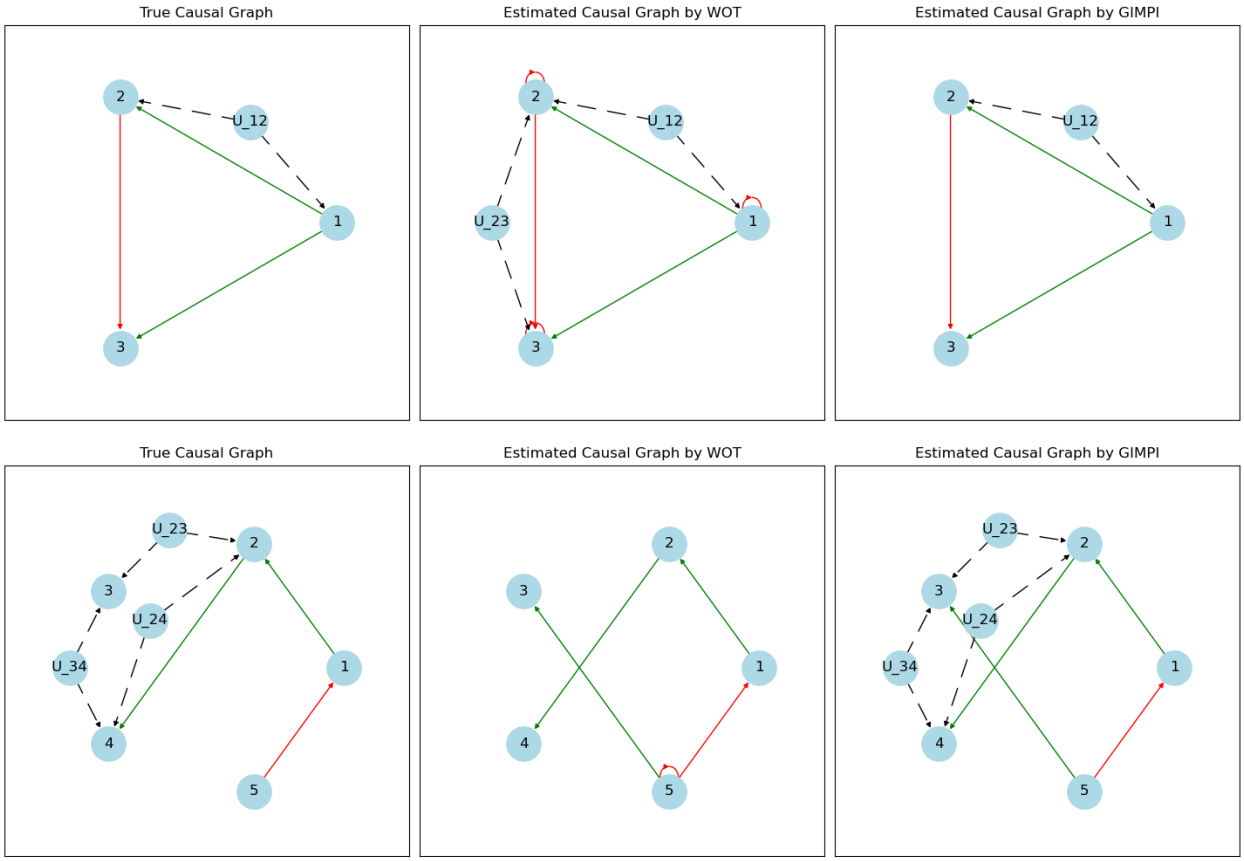


Figure 6: The true and estimated augmented causal graphs by WOT and APPEX for two random SDEs of dimensions $d = 3, 5$ are illustrated. Red edges $i \to j$ represent negative edge weights, such that $X^{(i)}$ negatively regulates $X^{(j)}$. Green edges $i \to j$ represent positive edge weights, such that $X^{(i)}$ positively regulates $X^{(j)}$. Dashed edges represent the effect of a latent confounder $U_{i,j}$ on observed variables i, j.

In terms of practical significance, we remark that identifiability is satisfied by a large class of distributions, including all nondegenerate discrete r.v's, and all nondegenerate continuous r.v's without elliptic symmetry (Exponential, Uniform, Gamma, etc.). Hence, one can easily check the initial distribution, or perform an intervention if needed, to guarantee system identifiability from data generated by a linear additive noise SDE. Furthermore, we have proven the optimality of each of APPEX's subprocedures, providing a solid justification for its practical use, as it improves its estimates with each iteration. This theoretical guarantee is reinforced by empirical evidence. Unlike WOT, APPEX accurately estimates drift and diffusion across hundreds of randomly generated SDEs between dimensions 2 to 10, in addition to resolving the classical non-identifiable systems from the literature, once provided with an appropriate initialization. APPEX therefore offers significant inference capability and flexibility compared to previous methods, since it does not rely on any knowledge of the drift or diffusion, in order to fully learn the dynamical system. 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. For example, APPEX can learn any linear drift regardless of the structure of the additive noise, whereas existing methods are strictly adapted to isotropic Brownian motion noise with a known diffusivity scale.

Although this paper presents several breakthroughs in the identification of SDEs from marginals, some limitations remain, offering opportunities for future research. First, we have restricted this work to linear drift and additive noise. Relaxing one or more of these assumptions, as well as time inhomogeneity, would certainly increase the practicality of our methods and theory. We note that more complex noise models, such as multiplicative noise, introduce additional sources of identifiability, such that the same marginal observations can be explained by a nonlinear additive noise SDE and a linear multiplicative noise SDE [CHS22]. Second, although we have shown that any initialization, which is not auto-rotationally invariant, will guarantee identifiability of linear additive noise SDEs, we have not studied their relative effectiveness towards parameter estimation. We conjecture that initializations, which are more drastically auto-rotationally asymmetric would be more robust for parameter estimation of arbitrary linear additive noise SDEs. Third, as shown in [Led09], the ML estimator of the AR(1) process (the discrete analogue of an Ornstein-Uhlenbeck process) is slightly biased. In future work, it would interesting to see if the MLEs we derived also provide biased parameter estimates, and if so, investigate possible improvements. Fourth, since we defined our method with respect to strict marginal constraints, APPEX is data hungry, and we have not yet developed precise asymptotics to characterize the rate of convergence given the number of iterations, amount of data per marginal, and time granularity. Combining our method with previous methods that relax these marginal constraints (see [LZKS21]), could improve estimates in low data settings, particularly if the measured marginals destroy samples. Finally, although we experimentally show that our method obeys consistency for randomly generated SDEs in the appendix, a formal proof of convergence and consistency would further strengthen its theoretical foundation.

To reiterate, in this paper we proved that the full identification of arbitrary linear additive noise SDEs from marginal snapshots is often possible, and we introduced a practical method to identify such systems. We believe these theoretical and methodological contributions will lead to breakthroughs across various domains, such as biology and hydrology, by enabling full dynamical system inference from population-level observational data.

Code Availability

The python code for reproducing the experimental results and figures is available soon. Analogous code in R is available at `https://github.com/HydroML/XOisAllYouNeed`.

Acknowledgements

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

References

[AG92] E Eric Adams and Lynn W Gelhar. Field study of dispersion in a heterogeneous aquifer: 2. spatial moments analysis. *Water Resources Research*, 28(12):3293–3307, 1992.

[ATW+24] Lazar Atanackovic, Alexander Tong, Bo Wang, Leo J Lee, Yoshua Bengio, and Jason S Hartford. Dyngfn: Towards bayesian inference of gene regulatory networks with gflownets. *Advances in Neural Information Processing Systems*, 36, 2024.

[AVI+20] Atte Aalto, Lauri Viitasaari, Pauliina Ilmonen, Laurent Mombaerts, and Jorge Gonçalves. Gene regulatory network inference from sparsely sampled noisy data. *Nature communications*, 11(1):3493, 2020.

[BCDMN19] Jean-David Benamou, Guillaume Carlier, Simone Di Marino, and Luca Nenna. An entropy minimization approach to second-order variational mean-field games. *Mathematical Models and Methods in Applied Sciences*, 29(08):1553–1583, 2019.

[BHCK23] Charlotte Bunne, Ya-Ping Hsieh, Marco Cuturi, and Andreas Krause. The schrödinger bridge between gaussian measures has a closed form. In *International Conference on Artificial Intelligence and Statistics*, pages 5802–5833. PMLR, 2023.

[BHR93] Keith J Beven, D Ed Henderson, and Alison D Reeves. Dispersion parameters for undisturbed partially saturated soil. *Journal of hydrology*, 143(1-2):19–43, 1993.

[Bis07] Jaya PN Bishwal. *Parameter estimation in stochastic differential equations*. Springer, 2007.

[BLL+20] Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

[BM24] Philip Boeken and Joris M Mooij. Dynamic structural causal models. *arXiv preprint arXiv:2406.01161*, 2024.

[BPRF06] Alexandros Beskos, Omiros Papaspiliopoulos, Gareth O Roberts, and Paul Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):333–382, 2006.

[BYB+92] J Mark Boggs, Steven C Young, Lisa M Beard, Lynn W Gelhar, Kenneth R Rehfeldt, and E Eric Adams. Field study of dispersion in a heterogeneous aquifer: 1. overview and site description. *Water Resources Research*, 28(12):3281–3291, 1992.

[CBL+19] Aaron G Cahill, Roger Beckie, Bethany Ladd, Elyse Sandl, Maximillian Goetz, Jessie Chao, Julia Soares, Cara Manning, Chitra Chopra, Niko 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.

[CCMV99] DW Chen, RF 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.

[CHLS23] Peter Craigmile, Radu Herbei, Ge Liu, and Grant Schneider. Statistical inference for stochastic differential equations. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(2):e1585, 2023.

[CHS22] Megan A Coomer, Lucy Ham, and Michael PH Stumpf. Noise distorts the epigenetic landscape and shapes cell-fate decisions. *Cell Systems*, 13(1):83–102, 2022.

[CML09] Paramita Chakraborty, Mark M Meerschaert, and Chae Young Lim. Parameter estimation for fractional transport: A particle-tracking approach. *Water resources research*, 45(10), 2009.

[CZHS22] Lénaïc Chizat, Stephen Zhang, Matthieu Heitz, and Geoffrey Schiebinger. Trajectory inference via mean-field langevin in path space. *Advances in Neural Information Processing Systems*, 35:16731–16742, 2022.

[DAD05] Frédérick Delay, Philippe Ackerer, and Charles Danquigny. Simulating solute transport in porous or fractured formations using random walk particle tracking: A review. *Vadose Zone Journal*, 4(2):360–379, 2005.

[Did08] Vanessa Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):245–264, 2008.

[Doo42] Joseph L Doob. The brownian movement and stochastic equations. *Annals of Mathematics*, 43(2):351–369, 1942.

[Elf06] Amro MM 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.

[FMR06] EO Frind, JW Molson, and DL Rudolph. Well vulnerability: a quantitative approach for source water protection. *Groundwater*, 44(5):732–742, 2006.

[For24] Aden Forrow. Consistent diffusion matrix estimation from population time series. *arXiv preprint arXiv:2408.14408*, 2024.

[Gil00] Daniel T Gillespie. The chemical langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000.

[HC15] Miranda Holmes-Cerfon. Applied stochastic analysis, 2015.

[HGJ16] Tatsunori Hashimoto, David Gifford, and Tommi Jaakkola. Learning population-level diffusions with generative rnns. In *International Conference on Machine Learning*, pages 2417–2426. PMLR, 2016.

[HS14] Niels Hansen and Alexander Sokol. Causal interpretation of stochastic differential equations. 2014.

[HSKP08] Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

[Hua24] Hanwen Huang. One-step data-driven generative model via schr\" odinger bridge. *arXiv preprint arXiv:2405.12453*, 2024.

[HZL+23] Mónica Basilio Hazas, Francesca Ziliotto, Jonghyun Lee, Massimo Rolle, and Gabriele 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.

[Jan21] Hicham Janati. *Advances in Optimal transport and applications to neuroscience*. PhD thesis, Institut Polytechnique de Paris, 2021.

[KOLL12] SC Kou, Benjamin P Olding, Martin Lysy, and Jun S Liu. A multiresolution method for parameter estimation of diffusion processes. *Journal of the American Statistical Association*, 107(500):1558–1574, 2012.

[LBSV+19] Luca Locatelli, Philip J Binning, Xavier Sanchez-Vila, Gitte Lemming Søndergaard, Louise Rosenberg, and Poul 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.

[Led09] Johannes 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.

[LES03] Russell Lande, Steinar Engen, and Bernt-Erik Saether. *Stochastic population dynamics in ecology and conservation*. Oxford University Press, USA, 2003.

[LJP+21] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3):218–229, 2021.

[LKR02] Peter C Lichtner, Sharad Kelkar, and Bruce Robinson. New form of dispersion tensor for axisymmetric porous media with implementation in particle tracking. *Water Resources Research*, 38(8):21–1, 2002.

[Löb14] Jörg-Uwe Löbus. Absolute continuity under time shift for ornstein-uhlenbeck type processes with delay or anticipation. *arXiv preprint arXiv:1411.7688*, 2014.

[LRW21] Yiheng Liu, Elina Robeva, and Huanqing Wang. Learning linear non-gaussian graphical models with multidirected edges. *Journal of Causal Inference*, 9(1):250–263, 2021.

[LZKS21] Hugo Lavenant, Stephen Zhang, Young-Heon Kim, and Geoffrey Schiebinger. Towards a mathematical theory of trajectory inference. *arXiv preprint arXiv:2102.09204*, 2021.

[Mad12] Kristóf Madarász. Information projection: Model and applications. *The Review of Economic Studies*, 79(3):961–985, 2012.

[MCF+24] Georg Manten, Cecilia Casolo, Emilio Ferrucci, Søren Wengel Mogensen, Cristopher Salvi, and Niki Kilbertus. Signature kernel conditional independence tests in causal discovery for stochastic processes. *arXiv preprint arXiv:2402.18477*, 2024.

[MFRC86] DM Mackay, DL Freyberg, PV Roberts, and JA 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.

[MH20] Søren Wengel Mogensen and Niels Richard Hansen. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1):539–559, 2020.

[MH22] Søren Wengel Mogensen and Niels Richard Hansen. Graphical modeling of stochastic processes driven by correlated noise. *Bernoulli*, 28(4):3023–3050, 2022.

[MHB16] Stephan Mandt, Matthew Hoffman, and David Blei. A variational analysis of stochastic gradient algorithms. In *International conference on machine learning*, pages 354–363. PMLR, 2016.

[MMP15] Kevin McGoff, Sayan Mukherjee, and Natesh Pillai. Statistical inference for dynamical systems: A review. 2015.

[MT07] Chuanjian Man and Christina 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.

[NM] Ahmed Nafidi and İlyasse Makroz. A comparison of methods for estimating parameters of the stochastic lomax process: through simulation study. *Hacettepe Journal of Mathematics and Statistics*, 53(2):495–505.

[NMY00] Jan Nygaard Nielsen, Henrik Madsen, and Peter C Young. Parameter estimation in stochastic differential equations: an overview. *Annual Reviews in Control*, 24:83–94, 2000.

[NR20] Richard Nickl and Kolyan Ray. Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *The Annals of Statistics*, 48(3):1383–1408, 2020.

[Nut21] Marcel Nutz. Introduction to entropic optimal transport. *Lecture notes, Columbia University*, 2021.

[O+02] Dennis R O'Connor et al. Part two: Report of the walkerton inquiry: A strategy for safe drinking water. 2002.

[Oks13] Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[OT10] Jungsun Oh and Christina W Tsai. A stochastic jump diffusion particle-tracking model (sjd-ptm) for sediment transport in open channel flows. *Water Resources Research*, 46(10), 2010.

[Özd16] Mustafa Özdemir. An alternative approach to elliptical motion. *Advances in Applied Clifford Algebras*, 26:279–304, 2016.

[Pau97] Anthony 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.

[Pav14] Grigorios A Pavliotis. Stochastic processes and applications. *Texts in Applied Mathematics*, 60, 2014.

[PC19] Gabriel Peyre and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[PHR19] Rich Pawlowicz, Charles Hannah, and Andy Rosenberger. Lagrangian observations of estuarine residence times, dispersion, and trapping in the salish sea. *Estuarine, Coastal and Shelf Science*, 225:106246, 2019.

[PMA03] 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.

[PP+08] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[RCM+24] Martin Rohbeck, Brian Clarke, Katharina Mikulik, Alexandra Pettet, Oliver Stegle, and Kai Ueltzhöffer. Bicycle: Intervention-based causal discovery with cycles. In *Causal Learning and Reasoning*, pages 209–242. PMLR, 2024.

[RSW21] Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

[Run21] Jakob 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.

[SBB24] Yunyi Shen, Renato Berlinghieri, and Tamara Broderick. Multi-marginal schr\" odinger bridges with iterative reference. *arXiv preprint arXiv:2408.06277*, 2024.

[SFGGH07] Peter Salamon, Daniel Fernàndez-Garcia, and JJ Gómez-Hernández. Modeling tracer transport at the made site: The importance of heterogeneity. *Water Resources Research*, 43(8), 2007.

[SHH+06] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[SST+19] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[Str08] Daniel W Stroock. *Partial differential equations for probabalists [sic]*. Number 112. Cambridge University Press, 2008.

[TB13] Shahab Torkamani and Eric A Butcher. Stochastic parameter estimation in nonlinear time-delayed vibratory systems with distributed delay. *Journal of Sound and Vibration*, 332(14):3404–3418, 2013.

[TLBB+23] Alejandro Tejada-Lapuerta, Paul Bertin, Stefan Bauer, Hananeh Aliee, Yoshua Bengio, and Fabian J Theis. Causal machine learning for single-cell genomics. *arXiv preprint arXiv:2310.14935*, 2023.

[VTLL21] Francisco Vargas, Pierre Thodoroff, Austen Lamacraft, and Neil Lawrence. Solving schrödinger bridges via maximum likelihood. *Entropy*, 23(9):1134, 2021.

[WGH+24] Yuanyuan Wang, Xi Geng, Wei Huang, Biwei Huang, and Mingming Gong. Generator identification for linear sdes with additive and multiplicative noise. *Advances in Neural Information Processing Systems*, 36, 2024.

[WTW+23] Lingfei Wang, Nikolaos Trasanidis, Ting Wu, Guanlan Dong, Michael Hu, Daniel E Bauer, and Luca Pinello. Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. *Nature Methods*, 20(9):1368–1378, 2023.

[WWT+18] Caleb Weinreb, Samuel Wolock, Betsabeh K Tusi, Merav Socolovsky, and Allon M Klein. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences*, 115(10):E2467–E2476, 2018.

[YWI+23] Toshiaki Yachimura, Hanbo Wang, Yusuke Imoto, Momoko Yoshida, Sohei Tasaki, Yoji Kojima, Yukihiro Yabuta, Mitinori Saitou, and Yasuaki Hiraoka. scegot: Single-cell trajectory inference framework based on entropic gaussian mixture optimal transport. *bioRxiv*, pages 2023–09, 2023.

[Zha24] Stephen Y Zhang. Joint trajectory and network inference via reference fitting. *arXiv preprint arXiv:2409.06879*, 2024.

[ZLSS24] Wenjun Zhao, Erica Larschan, Bjorn Sandstede, and Ritambhara Singh. Optimal transport reveals dynamic gene regulatory networks via gene velocity estimation. *bioRxiv*, pages 2024–09, 2024.

A Appendix

A.1 SDEs and causality

Let X_t be a stochastic process, with $V = [d]$ representing the set of endogenous variables $\{X_t^{(j)}\}_{j=1}^d$, and W representing the set of exogeneous variables, which are excluded from the principal model, but may influence the endogenous variables. A general SDE can be written in "integrand-integrator" form

$$dX_t = a(t, X_t)dh(t, X_t). \tag{40}$$

By integrating and considering each variable $X_t^{(j)}$ in the endogenous set $j \in V$, [BM24] showed that we can define a dynamic structural causal model (DSCM) [Definition 1]:

$$X_t^{(j)} = X_0^{(j)} + \int_0^t a_j(s, X_s^{\alpha(j)})dh_j(s, X_s^{\beta(j)}), \tag{41}$$

where $\alpha(j), \beta(j) \subset V \cup W$ represent the variables that are used as arguments for $a(t, X_t)$ and $h(t, X_t)$ respectively. In particular, (41) defines a causal graph $\mathcal{G}$ with vertices $V = [d]$, representing $X^{(1)}, \cdots X^{(d)}$, and edges $E = \{i \to j : j \in V, i \in \alpha(j) \cup \beta(j) \setminus \{j\}\}$ [BM24][Definition 2]. Intuitively, we include the edge $i \to j$ if $X^{(i)}$ influences the evolution of variable $X^{(j)}$ through the integrand $a_j(t, X_t)$ or integrator $h_j(t, X_t)$. For simplicity, self-edges $i \to i$ can be omitted [BM24].

For instance, any time homogeneous additive noise SDE driven by Brownian motion

$$dX_t = b(X_t)dt + GdW_t, \tag{42}$$

can be rewritten in the form (40) by setting $a(t, X_t) = \begin{bmatrix} b(X_t) & G \end{bmatrix}$ and $h(t, X_t) = \begin{bmatrix} t & W_t \end{bmatrix}^T$. Given that $X_t \in \mathbb{R}^d$, $W_t \in \mathbb{R}^m$, we have corresponding dimensions $b(X_t) \in \mathbb{R}^d$, $\sigma(X_t) \in \mathbb{R}^{d \times m}$, $Z_t \in \mathbb{R}^{m+1}$, and $a(X_t) \in \mathbb{R}^{d \times (m+1)}$. This admits the DSCM:

$$X_t^{(j)} = X_0^{(j)} + \int_0^t b_j(X_s^{\alpha(j)})ds + \int_0^t G_j dW_s. \tag{43}$$

Under the assumption of independent driving noise, also known as independent integrators [BM24][Theorem 4], $\beta(j) \subset W$ and $\beta(i) \cap \beta(j) = \emptyset$ for all $i, j \in V$, DSCMs observe graphical Markov properties under local independence [Did08, MH20, MH22]. Given the additive noise model (42), this condition is equivalent to causal sufficiency. Indeed, additive noise implies $\beta(j) \subset W$, which rules out instantaneous effects between endogenous variables, and $\beta(i) \cap \beta(j) = \emptyset$ holds if and only if there are no unobserved confounders.

Remark 2. *Most works studying continuous stochastic processes consider independent driving noise. Correlated driving noise was first studied in detail in [MH22], and emerges from variables sharing driving noise sources (see Example 4 in [BM24]). Correlated driving noise crucially allows the causal model to capture unobserved confounders via bidirected or multidirected edges in the causal graph, whereas independent driving noise implicitly assumes causal sufficiency.*

In the case of an additive noise process with diffusion $H = GG^T$, we see that the driving Brownian noise is independent among components if and only if $G_i \cdot G_j = 0$ for all $i \neq j$. This is due to multivariate Brownian motion being jointly Gaussian across its components, and the property that jointly Gaussian distributions are independent if and only if they are uncorrelated. In contrast, we can model correlated driving noise if $G_i \cdot G_j \neq 0$ for some $i \neq j$. Indeed, $H_{ij} \neq 0$ would make the variables $X_t^{(i)}$ and $X_t^{(j)}$ share a common noise source $dW_t^{(k)}$.

Example 1. *Consider two matrices G_1, G_2, which share the same observational diffusion H:*

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

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 common noise source that is shared among all 3 variables, since each column contains a 0. Hence, the causal interpretation under G_1 consists in three bidirected edges $1 \leftrightarrow 2$, $1 \leftrightarrow 3$, $2 \leftrightarrow 3$. In contrast, the causal interpretation under G_2 consists in a single multidirected edge $(1, 2, 3)$, since all components share noise sources.

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 $[d] \setminus \{i, j\}$.

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 Theorem **??**, simple 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 $\sigma(X_t)\sigma(X_t)^T$ 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 1, where we saw one interpretation feature three bidirected edges and another feature a multi-edge, but under non-additive noise, the ambiguity extends to simple edges $i \to j$. This is illustrated in Example 5.5 in [HS14].

A.2 Weak formulation of Fokker-Planck

We may define the linear operator $\mathcal{L}$

$$(\mathcal{L}f)(x) = b(x) \cdot f(x) + \frac{H}{2} : \nabla^2 f(x),$$

where : denotes the Kronecker product. $\mathcal{L}$ is the generator of the additive noise SDE (1) [HC15]. Its adjoint operator is given by

$$(\mathcal{L}^* g)(x) = -\nabla \cdot (b(x,t)g(x)) + \frac{H}{2}\nabla^2 g(x).$$

and satisfies

$$\int_{\mathbb{R}^d} \mathcal{L}f(x)g(x)dx = \langle \mathcal{L}f, g \rangle = \langle f, \mathcal{L}^* g \rangle = \int_{\mathbb{R}^d} f(x)\mathcal{L}^* g(x)dx,$$

for all smooth and compactly supported test functions $f, g \in C_c^\infty(\mathbb{R}^d)$.

Then, the Fokker-Planck equation can be formulated in the weak sense as follows. For each $t \geq 0$ let ρ_t be a probability measure on $\mathbb{R}^d$ and assume that $t \mapsto \rho_t$ is narrowly continuous. Then we say ρ_t solves the Fokker-Planck equation in the weak sense provided that

$$\frac{d}{dt}\int \varphi(x)d\rho_t(x) = \int \mathcal{L}\varphi(x)d\rho_t(x) \qquad \forall \varphi \in C_c^2(\mathbb{R}^d).$$

This weak formulation accommodates the case where b is not differentiable and ρ_t is not necessarily absolutely continuous with respect to the Lebesgue measure. For ρ_t with sufficiently regular Lebesgue density $p(x,t)$ this weak solution concept is equivalent to the classical sense of the Fokker-Planck equation, i.e.

$$\partial_t p(x,t) = \mathcal{L}^* p(x,t).$$

For the measure-theoretic weak formulation of the Fokker-Planck equation, we quote the following existence and uniqueness results from [Str08]. Under our standing assumptions on the drift and diffusion coefficients (namely: the drift is linear, and the diffusion is constant but possibly degenerate), by [Str08] Theorem 1.1.9 weak solutions to the Fokker-Planck equation exist whenever the initial condition ρ_0 has finite second moments; and moreover, by Theorem 2.2.9 in [Str08], the solution is unique provided that all moments of ρ_0 are finite.

A.3 Generalized rotations and additional proofs

Definition 3. *A rotation in $\mathbb{R}^d$ is defined via the matrix exponential e^{At} where $A \in \mathbb{R}^{d\times d}$ is skew-symmetric, i.e. $A + A^T = 0$.*

Example 2. *The unit sphere in $\mathbb{R}^d$*

$$S^{d-1} = \{x \in \mathbb{R}^d : x^T x = 1\}$$

is rotationally invariant, i.e. $\forall t \geq 0$, $e^{At} S^{d-1} = S^{d-1}$ for a skew-symmetric matrix A, i.e. $A + A^T = 0$.

Proof. For any $x \in S^{d-1}$, we have $y = e^{At}x \in S^{d-1}$, since

$$y^T y = (e^{At}x)^T e^{At}x = x^T e^{A^T t} e^{At} x = x^T e^{(A+A^T)t} x = x^T x = 1.$$

Similarly, the inverse map $y = e^{-At}x \in S^{d-1}$. □

Example 3. *Let $\Sigma \succ 0$ and $r > 0$. An ellipsoid defined by the quadratic form*

$$E_\Sigma = \{x \in \mathbb{R}^d : x^T \Sigma^{-1} x = r\}$$

*is Σ-rotationally invariant, i.e. $\forall t \geq 0$, $e^{At} E_\Sigma = E_\Sigma$ if $A\Sigma + \Sigma A^T = 0$ (see Lemma **??**).*

Proof. For any $x \in E_\Sigma$, we have $y = e^{At}x \in E_\Sigma$, since

$$y^T \Sigma y = (e^{At}x)^T \Sigma e^{At} x = x^T (e^{A^T t} \Sigma e^{At}) x = x^T \Sigma x = 1.$$

Lemma 3 was applied in the penultimate step. Similarly, the inverse map $y = e^{-At}x \in E_\Sigma$. □

Lemma 1. *Let X be a r.v. with covariance Σ and suppose that $e^{At}X \sim X$. Then, $A\Sigma + \Sigma A^T = 0$.*

Proof. Let $C(t) = Cov(e^{At}X)$. $e^{At}X \sim X$ implies that for all $t \geq 0$, $C(t) = Cov(X) = \Sigma$. Hence,

$$\begin{aligned} C(t) &= e^{At}\Sigma e^{A^T t} = \Sigma \\ C'(t) &= Ae^{At}\Sigma e^{A^T t} + e^{At}\Sigma A^T e^{A^T t} = 0 \\ C'(0) &= A\Sigma + \Sigma A^T = 0. \end{aligned}$$

□

Lemma 2. *Let $Cov(X) = \Sigma \succ 0$ and suppose that X is auto-rotationally invariant. Then X admits a density, and the level sets of its probability density function are Σ-rotationally invariant ellipsoids:*

$$\text{Let } E_\Sigma^{(k)} = \{x \in \mathbb{R}^d : x^T \Sigma^{-1} x = k\}, \text{ then } \forall x, y \in E_\Sigma^{(k)}, p(x) = p(y) \tag{44}$$

Proof. If X is auto-rotationally invariant, then there is a Σ-generalized rotation $e^{At}X \sim X$ with $A \neq 0$. Equivalently, $p(x) = p(e^{At}x)$ for all $x \in \mathbb{R}^d$. Since $e^{At}E_\Sigma^{(k)} = E_\Sigma^{(k)}$, the probability distribution of X must be constant within each ellipsoid $E_\Sigma^{(k)}$. Since $\Sigma \succ 0$, then each set $E_\Sigma^{(k)}$ is non-degenerate in $\mathbb{R}^d$, thus inducing a probability density function on X. □

Lemma 3. *Given $A\Sigma + \Sigma A^T = 0$, it follows that*

$$(e^{At})^T \Sigma e^{At} = \Sigma.$$

Proof. Let $M(t) = (e^{At})^T \Sigma e^{At}$. Then,

$$\begin{aligned}\frac{d}{dt}M(t) &= (Ae^{At})^T \Sigma e^{At} + (e^{At})^T \Sigma (Ae^{At}) \\ &= e^{A^T t}(A^T\Sigma + \Sigma A)e^{At} = 0,\end{aligned}$$

where we used $A\Sigma + \Sigma A^T = 0$ in the last step. Hence, $M(t) = (e^{At})^T \Sigma e^{At} = M(0) = \Sigma$ □

Lemma 4. *Let X be a d dimensional r.v. such that $dim(span(X)) < d$ with probability 1. Then,*

$$e^{At}X \sim X$$

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

Proof. We expand

$$e^{At}X = X + \left(\sum_{n=1}^{\infty} \frac{t^n}{n!}A^n\right)X$$

Hence, a sufficient condition for the claim is that X is in the nullspace of A with probability 1. Since $dim(span(X)) < d$ a.s., then by the rank-nullity theorem, there exists A such that $AX = 0$ a.s. □

Lemma 5. *Let $\Sigma \succeq 0$ be a symmetric positive semidefinite matrix and suppose that A is skew-symmetric with respect to Σ, i.e.*

$$A\Sigma + \Sigma A^T = 0,$$

then without loss of generality, we may define A with $Tr(A) = 0$.

Proof.

$$\mathrm{Tr}(A\Sigma + \Sigma A^T) = \mathrm{Tr}(0) = 0.$$

Using the linearity of the trace and the cyclic property $\mathrm{Tr}(XY) = \mathrm{Tr}(YX)$:

$$\begin{aligned}\mathrm{Tr}(A\Sigma + \Sigma A^T) &= \mathrm{Tr}(A\Sigma) + \mathrm{Tr}(A^T\Sigma) \\ &= \mathrm{Tr}(A\Sigma) + \mathrm{Tr}(\Sigma A) \\ &= 2\mathrm{Tr}(A\Sigma).\end{aligned}$$

Therefore,

$$2\mathrm{Tr}(A\Sigma) = 0 \quad \Rightarrow \quad \mathrm{Tr}(A\Sigma) = 0.$$

If $\Sigma \succ 0$, then we immediately conclude $Tr(A) = 0$. Otherwise, we may pick A such that $\Sigma_{i,i} = 0 \implies A_{i,i} = 0$ to ensure that $Tr(A) = 0$.

□

Proposition 4. *Let $X_0 \sim p_0$ be a discrete random variable. Then X_0 is not auto-rotationally invariant if and only if the support of p_0 spans $\mathbb{R}^d$.*

Proof. Lemma 4 provides the only if direction.

Now, let $span(X_0) = \mathbb{R}^d$ a.s. Since e^{At} represents a continuous transformation, and X_0 is supported on a discrete set of points, which cannot all be mapped to 0 by A, it follows that the support would shift for each $t > 0$. We conclude that $e^{At}X_0 \not\sim X_0$. □

Lemma 6. *As in Theorem 1, let X_0 have covariance $\Sigma = GG^T$ be auto-rotationally invariant, such that $e^{As}X_0 \sim X_0\ \forall s \geq 0$, and consider the SDEs*

$$dX_t = \gamma G dW_t \tag{45}$$
$$dX_t = AX_t dt + \gamma G dW_t. \tag{46}$$

Then, X_t will also be auto-rotationally invariant, such that $e^{As}X_t \sim X_t\ \forall s \geq 0$.

Proof. Since both SDEs are linear additive noise, they admit closed form solutions. For the first SDE (45), we have

$$X_t = X_0 + \gamma \int_0^t G dW_s.$$

Moreover, because the SDE parameters A, G obey $GG^T = \Sigma$ and $A\Sigma + \Sigma A^T = 0$ by construction, then by the Σ-rotational invariance of $\mathcal{N}(0, \Sigma)$, we have

$$G dW_s \sim \mathcal{N}(0, \Sigma ds) \sim e^{A(t-s)} G dW_s.$$

We may therefore write the solution of the first SDE (45) as

$$X_t = X_0 + \gamma \int_0^t e^{A(t-s)} G dW_s.$$

Hence, $X_t \sim X_0 + M_t$, where $M_t = \gamma \int_0^t e^{A(t-s)} G dW_s \sim \mathcal{N}(0, \gamma^2 \Sigma)$. Since X_0 and M_t are each auto-rotationally invariant with respect to the map e^{As}, while being independent due to the independent increments of Brownian motion, it follows that $X_t \sim X_0 + M_t$ is also auto-rotationally invariant with respect to the map e^{As}. We immediately recognize the same conclusion for the second SDE (46), since its closed solution is the same in distribution to the first SDE:

$$X_t = e^{At}X_0 + \gamma \int_0^t e^{A(t-s)} G dW_s \sim X_0 + \gamma \int_0^t e^{A(t-s)} G dW_s.$$

□

The linear additive noise SDE can be generalized to accommodate a constant drift μ:

$$dX_t = (AX_t + \mu)\, dt + G dW_t, X_0 \sim p_0 \tag{47}$$

Corollary 1. *Let the initial distribution p_0 have finite support on $d+1$ vectors $\{x_i : x_i \in \mathbb{R}^d\}_{i=1}^{d+1}$ spanning $\mathbb{R}^d$ such that each pair of vectors are linearly independent. If X_t evolves according to an arbitrary d-dimensional linear additive noise SDE with an additional constant drift μ: (47), then μ, A, and GG^T of the SDE are uniquely identified from temporal marginals.*

Proof. Following the same argument as Theorem (1), we obtain a deterministic residual SDE with respect to $X_0 \sim p(x, 0)$:

$$\frac{dX_t}{dt} = \bar{A}X_t + \bar{\mu} = 0.$$

This implies that $\bar{A}x_i = -\bar{\mu}$ for each of the $d+1$ vectors x_i in the support of p_0. Equivalently, since each pair of vectors is linearly independent, we can write

$$\bar{A}(x_i - x_j) = 0,$$

such that the differences $x_i - x_j$ span $\mathbb{R}^d$. The rank-nullity theorem implies that $\bar{A} = 0$ and hence $\bar{\mu} = 0$. □

A.4 Revisiting non-identifiablity examples

The proof of Theorem (1) provides a nice characterization of non-identifiability. A pair of SDEs are non-identifiable with respect to one another given $p(x,0)$ only if $p(x,0)$ is a stationary distribution with respect to their residual SDE.

Pair 1: Starting at stationary
In this example, the residual SDE is given by

$$dX_t = -9X_t dt + \sqrt{9}dW_t,$$

for which $\mathcal{N}(0, \frac{1}{2})$ is a stationary distribution, confirming non-identifiability from $p_0 \sim \mathcal{N}(0, \frac{1}{2})$. However, it is obvious that a discrete probability distribution cannot be stationary with respect to the residual SDE. Intuitively, the non-identifiability breaks down when we start away from the two SDEs' shared stationary distribution, since the SDEs can then be identified by different rates of convergence to the stationary distribution, as shown in Figure 2a

Pair 2: Rotation around process mean
In this example, the residual SDE is given by

$$\frac{dX_t}{dt} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} X_t \implies X_t = \begin{bmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{bmatrix} X_0.$$

We can therefore confirm non-identifiability if $X_0 \sim \mathcal{N}(0, Id)$, since $\begin{bmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{bmatrix}$ corresponds to a counterclockwise rotation about the origin, and $\mathcal{N}(0, Id)$ is isotropic. Note that an initial distribution supported on just a single non-zero vector X_0 would suffice in identifying the rotation for this example since $dim(null(A)) = 0$. This is shown in Figure 2b for $X_0 = (2, 0)$.

Pair 3: Degenerate Rank In this example, the residual SDE is given by

$$\frac{dX_t}{dt} = \begin{bmatrix} 2/3 & 2/3 \\ 1/3 & 1/3 \end{bmatrix} X_t.$$

If $X_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ then X_0 is stationary with respect to the residual SDE since $dX_t = \begin{bmatrix} 2/3 & 2/3 \\ 1/3 & 1/3 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 0$, as is the case for any X_0 in $span(\begin{bmatrix} 1 \\ -1 \end{bmatrix})$. However, if p_0 is supported on two linearly independent points, then it would be impossible to jointly achieve $dX_t = 0$ since $dim(null(A)) = 1$. This is shown in Figure 2c. Intuitively, this initialization resolves the non-identifiablity by guaranteeing that both SDEs are observed on the entire space $\mathbb{R}^d$, whereas the previous initialization restricted observation to a linear subspace where the SDEs happened to behave equivalently.

Remark 3. *We have seen that p_0 being stationary with respect to the residual SDE is a necessary condition for a pair of SDEs to be non-identifiable from p_0. However, this condition is not sufficient. Two SDEs may be identifiable from p_0 despite p_0 being stationary with respect to the residual SDE. A simple example is given by a modification of the degenerate rank example.*

$$dX_t = \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix} X_t dt + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} dW_t, X_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{48}$$

$$dY_t = \begin{bmatrix} 1/3 & 4/3 \\ 2/3 & -1/3 \end{bmatrix} Y_t dt + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} dW_t, X_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{49}$$

This pair of SDEs will have the same residual SDE as in the unidentifiable example, which is stationary with respect to $X_0 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$. However, by adjusting the diffusion to be full rank, this ensures that the processes will diffuse away from $span(\begin{bmatrix} 1 \\ -1 \end{bmatrix})$, which would make the processes identifiable by their different behaviour away from this subspace.

A.5 Maximum Likelihood Estimation

Proof of Theorem (3). We follow the standard procedure [PMA03, Pav14] for deriving maximum likelihood estimators. 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 [Pav14], we consider the discretized law $X_{i+1}|X_i \sim \mathcal{N}(X_i + AX_i dt, Hdt)$, which implies

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

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

$$\mathcal{L} = \frac{1}{(2\pi dt)^{\frac{dM(N-1)}{2}} det(H)^{\frac{M(N-1)}{2}}} \exp\left(-\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}\frac{1}{2}(\Delta X_i^{(j)} - AX_i^{(j)}dt)^T (Hdt)^{-1}(\Delta X_i^{(j)} - AX_i^{(j)}dt)\right)$$

$$\log(\mathcal{L}) = -\frac{M(N-1)}{2}\left(d\log(2\pi dt) + \log(\det(H))\right) - \frac{1}{2}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}\left(\Delta X_i^{(j)} - AX_i^{(j)}dt)^T (Hdt)^{-1}(\Delta X_i^{(j)} - AX_i^{(j)}dt)\right)$$

We then derive the maximum likelihood estimators through matrix calculus [PP+08]:

$$\begin{aligned}
\frac{d\log(L)}{dA} &= -\frac{1}{2}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}\frac{d}{dA}\left(\Delta X_i^{(j)} - AX_i^{(j)}dt)^T (Hdt)^{-1}(\Delta X_i^{(j)} - AX_i^{(j)}dt)\right) \\
&= -\frac{1}{2}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1} -2dt\frac{d}{dA}\left((\Delta X_i^{(j)})^T (Hdt)^{-1} AX_i^{(j)}\right) + dt^2\frac{d}{dA}\left(X_i^{(j)^T} A^T (Hdt)^{-1} AX_i^{(j)}\right) \\
&= -\frac{1}{2}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1} -2dt(Hdt)^{-1}(\Delta X_i^{(j)})X_i^{(j)^T} + 2dt^2(Hdt)^{-1}AX_i^{(j)}X_i^{(j)^T} \\
&= \sum_{i=0}^{N-2}\sum_{j=0}^{M-1}(H)^{-1}(\Delta X_i^{(j)})X_i^{(j)^T} - dt(H)^{-1}AX_i^{(j)}X_i^{(j)^T}
\end{aligned}$$

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

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

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

$$\begin{aligned}
\frac{d\log(L)}{dH^{-1}} &= \frac{d}{dH^{-1}}\left(\frac{M(N-1)}{2}\log(\det(H^{-1})) - \frac{1}{2dt}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}\left(\Delta X_i^{(j)} - AX_i^{(j)}dt)^T H^{-1}(\Delta X_i^{(j)} - AX_i^{(j)}dt)\right)\right) \\
&= \frac{M(N-1)}{2}H - \frac{1}{2dt}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}\left(\Delta X_i^{(j)} - AX_i^{(j)}dt)(\Delta X_i^{(j)} - AX_i^{(j)}dt)^T\right)
\end{aligned}$$

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

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

□

Under a general additive noise SDE

Corollary 2. *Let X_t evolve according to a general additive noise SDE (1). Then the log-likelihood function is given by*

$$\mathcal{L} = \frac{1}{(2\pi dt)^{\frac{dM(N-1)}{2}}det(H)^{\frac{M(N-1)}{2}}}\exp\left(-\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}\frac{1}{2}(\Delta X_i^{(j)} - b(X_i^{(j)})dt)^T(Hdt)^{-1}(\Delta X_i^{(j)} - b(X_i^{(j)})dt)\right)$$
$$\log(\mathcal{L}) = -\frac{M(N-1)}{2}\left(d\log(2\pi dt) + \log(\det(H))\right) - \frac{1}{2}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}\left(\Delta X_i^{(j)} - b(X_i^{(j)})dt)^T(Hdt)^{-1}(\Delta X_i^{(j)} - b(X_i^{(j)})dt)\right)$$

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

$$0 = -\frac{1}{2}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}\frac{d}{d\alpha_b^{(k)}}\left(\Delta X_i^{(j)} - b(X_i^{(j)})dt)^T(Hdt)^{-1}(\Delta X_i^{(j)} - b(X_i^{(j)})dt)\right)$$

The MLE for the diffusion H admits the closed solution

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

Theorem 2 (Exact MLE estimators for drift and diffusion of 1 dimensional SDE (2) from multiple trajectories)**.** *Given M different trajectory time series over N different times: $\{X_{t_i}^{(j)} : i \in 0, ..., N-1,\ j \in 0, ..., M-1\}$ sampled from the linear additive noise SDE (2), the maximum likelihood solution for linear drift is approximated by*

$$\hat{A} = \frac{1}{dt}\log\left(\frac{\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}X_{i+1}^{(j)}X_i^{(j)}}{\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}X_i^{(j)^2}}\right) \tag{50}$$

and the maximum likelihood solution for diffusion is approximated by

$$\hat{\sigma}^2 = \frac{1}{M(N-1)dt}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}(X_{i+1}^{(j)} - e^{Adt}X_i^{(j)})^2 \tag{51}$$

Proof. We proceed as in the proof of Theorem (**??**). The exact log-likelihood for the one dimensional case is

$$\log(\mathcal{L}) = -\frac{M(N-1)}{2}\left(\log(2\pi dt) + \log(\sigma^2)\right) - \frac{1}{2\sigma^2 dt}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}(X_{i+1}^{(j)} - e^{Adt}X_i^{(j)})^2$$

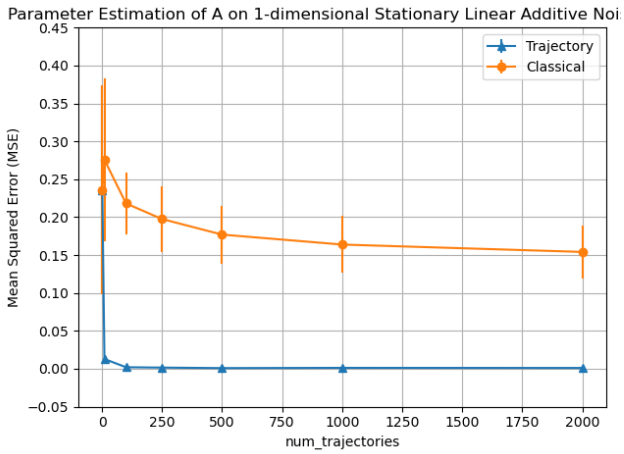


Figure 7: The expectation of the classical MLE drift estimator (Classical) converges at a slower rate compared to the MLE drift estimator (50) derived for multiple trajectories (Trajectory).

To solve for $\hat{A}$, we compute

$$0 = \frac{\partial \log(\mathcal{L})}{\partial A} \propto \sum_{i=0}^{N-2}\sum_{j=0}^{M-1} \frac{\partial}{\partial A}\left(-2X_{i+1}^{(j)}X_i^{(j)}e^{Adt} + e^{2Adt}{X_i^{(j)}}^2\right)$$

$$e^{2Adt}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}{X_i^{(j)}}^2 = e^{Adt}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}X_{i+1}^{(j)}X_i^{(j)}$$

$$e^{Adt} = \frac{\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}X_{i+1}^{(j)}X_i^{(j)}}{\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}{X_i^{(j)}}^2}$$

$$A = \frac{1}{dt}\log\left(\frac{\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}X_{i+1}^{(j)}X_i^{(j)}}{\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}{X_i^{(j)}}^2}\right)$$

Similarly, to solve for $\hat{\sigma}^2$, we compute

$$0 = \frac{\partial \log(\mathcal{L})}{\partial \sigma^2} \propto -\frac{M(N-1)}{2\sigma^2} + \frac{1}{2dt}\frac{1}{(\sigma^2)^2}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}(X_{i+1}^{(j)} - e^{Adt}X_i^{(j)})^2$$

$$\frac{M(N-1)}{2} = \frac{1}{2dt\sigma^2}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}(X_{i+1}^{(j)} - e^{Adt}X_i^{(j)})^2$$

$$\sigma^2 = \frac{1}{M(N-1)dt}\sum_{i=0}^{N-2}\sum_{j=0}^{M-1}(X_{i+1}^{(j)} - e^{Adt}X_i^{(j)})^2$$

□

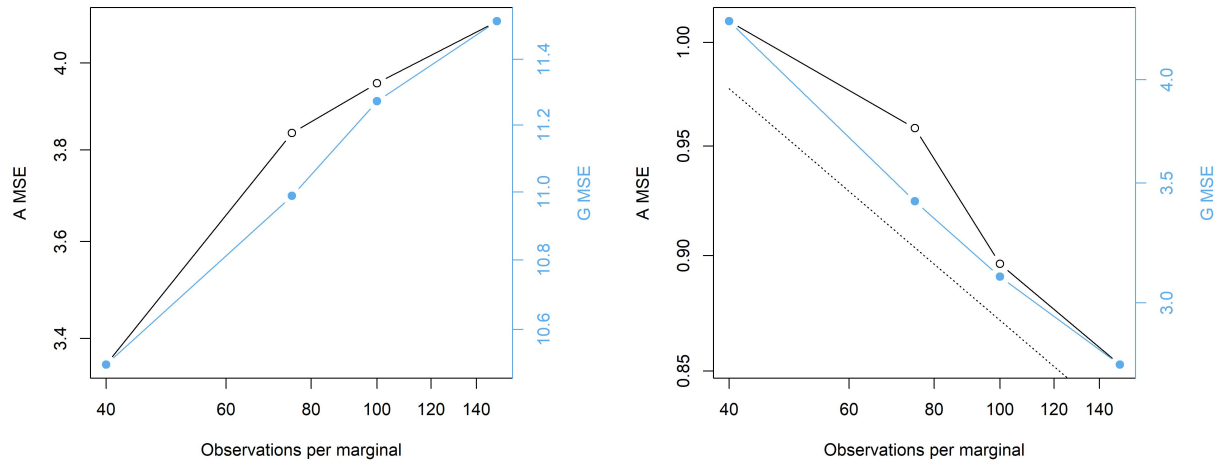
Remark 4. *Previous works have predominately focused on the case of observing a single long trajectory rather than a collection of short trajectories. Suppose that one observes a set of N observed trajectories, then drift estimation may also be performed by averaging the classical MLE estimator for a single trajectory across observations: $\mathbb{E}_N[\hat{A}_T] = \frac{1}{N}\frac{\int_0^T X_t dX_t}{\int_0^T X_t^2 dt}$. Note that this is distinct from the MLE estimator that we derived for multiple trajectories in (50). Indeed, we observe that the latter estimator converges at a much faster rate than the averaged classical estimator.*

We now derive maximum likelihood solutions for the

Theorem 3.

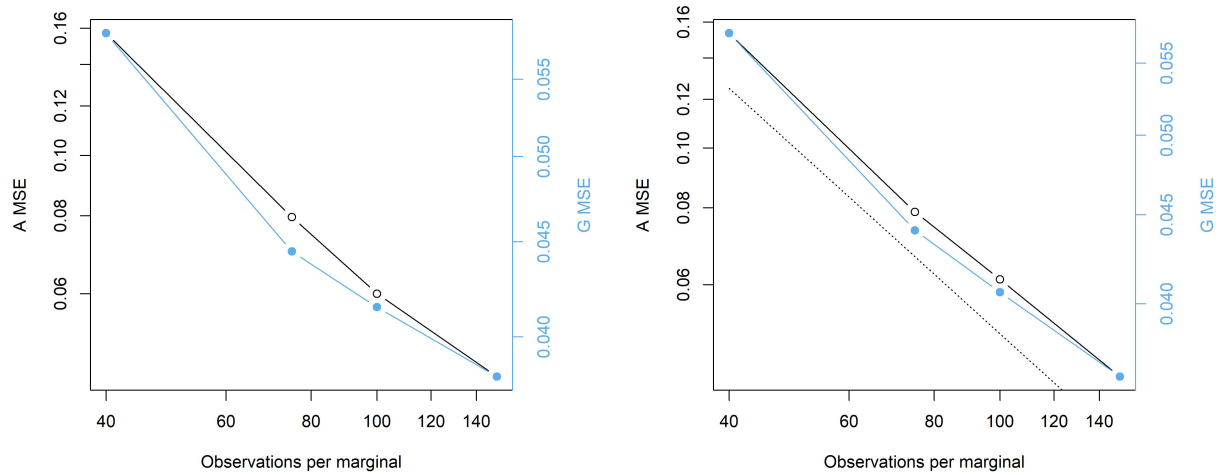
B Additional experiments

B.1 Consistency experiments



(a) WOT (b) GIMPI

Figure 8: Average mean squared error (MSE) of A and G as the number of observations per marginal increases across 50 random 2d systems. Note this is a log-log plot, and the dotted line in the GIMPI plot shows that the convergence follows a power law relationship with power -0.125.



(a) WOT (b) GIMPI

Figure 9: Average mean squared error (MSE) of A and G as the number of observations per marginal increases across 70 random 12d systems. Note this is a log-log plot, and the dotted line in the GIMPI plot shows that the convergence follows a power law relationship with power -1.