Efficient Learning of Stationary Diffusions with Stein-type Discrepancies

Fabian Bleile
Technical University of Munich
Munich Center for Machine Learning

Sarah Lumpp
Technical University of Munich

Mathias Drton
Technical University of Munich
Munich Center for Machine Learning

Abstract

Learning a stationary diffusion amounts to estimating the parameters of a stochastic differential equation whose stationary distribution matches a target distribution. We build on the recently introduced kernel deviation from stationarity (KDS), which enforces stationarity by evaluating expectations of the diffusion's generator in a reproducing kernel Hilbert space. Leveraging the connection between KDS and Stein discrepancies, we introduce the Stein-type KDS (SKDS) as an alternative formulation. We prove that a vanishing SKDS guarantees alignment of the learned diffusion's stationary distribution with the target. Furthermore, under broad parametrizations, SKDS is convex with an empirical version that is ϵ-quasiconvex with high probability. Empirically, learning with SKDS attains comparable accuracy to KDS while substantially reducing computational cost, and yields improvements over the majority of competitive baselines.

1 INTRODUCTION

Understanding cause-and-effect relationships is fundamental across many scientific disciplines, from the social sciences to natural sciences and engineering. Causal modeling provides a principled framework for reasoning about such relationships, enabling predictions under interventions and guiding decision-making in complex systems. Traditional approaches to causal inference, often built on structural causal models (SCMs), assume an underlying directed acyclic graph (DAG) that encodes causal dependencies. While this assumption facilitates theoretical analysis and algorithmic implementation, it poses challenges in systems with cyclic dependencies (Spirtes et al., 2000; Pearl, 2009).

A causal model is designed to represent observational, interventional, and sometimes counterfactual distributions. Stationary Itô diffusions, modeled through stochastic differential equations (SDEs), provide a natural framework for representing causal relationships in systems where cyclic dependencies arise (Lorch et al., 2024; Varando and Hansen, 2020; Sokol and Hansen, 2014). Unlike traditional SCMs, this approach inherently accommodates cyclic dependencies while ensuring a well-defined stationary distribution that characterizes long-term system behavior. In the causal framework of stationary diffusions, the observational distribution of a d-dimensional random variable X is modeled as the stationary distribution P of a stochastic process $(X_t)_{t\geq 0}$, defined by the following SDE

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t, \tag{1}$$

with drift function b and diffusion function σ (Fig. 1). Intuitively, the drift term b describes the deterministic tendency of the process, while the diffusion term σ models random fluctuations driven by Brownian motion. The distribution P is called stationary if $(X_t)_{t\geq 0}$ solves (1) with $X_t \sim P$ for every $t \geq 0$. Under mild conditions ensuring that the process is stable and does not diverge, such a stationary solution exists and is unique (Huang et al., 2015).

Let p be a probability density, then the Langevin diffusion $dX_t = \nabla \log p(X_t)dt + \sqrt{2}I_d dB_t$ is a stationary diffusion with stationary density p. This reduces the task of finding a causal model, as described, to a score estimation problem. However it is non-trivial how to incorporate interventional distributions into this model. Interventions in this causal framework are modeled by modifying the drift and diffusion functions. The interventional distribution is then, assuming existence and uniqueness, the respective stationary distribution

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

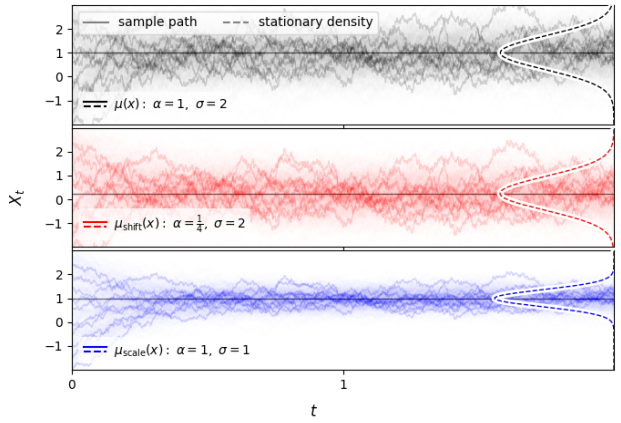

Figure 1: Sample paths and stationary densities of the one-dimensional Ornstein–Uhlenbeck process $dX_t = -4(X_t - \alpha)dt + \sigma dB_t$. The top panel shows the ground-truth model ($\alpha = 1, \sigma = 2$), while the middle and bottom panels illustrate a shifted mean ($\alpha = \frac{1}{4}, \sigma = 2$) and a changed scale ($\alpha = 1, \sigma = 1$). Dashed curves indicate the corresponding stationary densities.

(Sokol and Hansen, 2014; Lorch et al., 2024). In contrast to traditional SCMs, which refer to structural equations among the observed variables directly, the stationary diffusion causal framework focuses on joint distributions that emerge from a (possibly intervened) system's equilibrium behavior.

Consider a model for the SDE in (1) that treats drift and diffusion terms as unknown parameters. Learning a stationary diffusion means adjusting its parameters so that a stationary solution P exists and matches an observed target distribution μ. Lorch et al. (2024) propose a loss—*Kernel Deviation from Stationarity* (KDS)—that measures the discrepancy between μ and P using only samples of μ and the parameters of b and σ. This enables direct optimization of the SDE parameters. This approach performs well in small simulations but is computationally expensive and, thus, not feasible for data of larger dimension. Our main contribution is a novel loss function for learning stationary SDEs—the *Stein-type Kernel Deviation from Stationarity (SKDS)*, which offers significantly lower computational cost at similar estimation performance. We motivate SKDS from Stein's method, inspired by the work on Stein discrepancies (Gorham and Mackey, 2015; Liu et al., 2016; Gorham and Mackey, 2017), and show its practical and theoretical advantages over KDS in learning from interventional data. Before introducing SKDS, we review relevant results from kernel theory, Stein discrepancies, and the martingale characterization of stationarity.

2 BACKGROUND

We denote the space of functions from a domain U to a vector space V by $\Gamma(U, V)$; and we write $\Gamma(U) := \Gamma(U, \mathbb{R})$ for real-valued functions. We define $C^n(U, V)$ (respectively, $C_b^n(U, V)$) as the set of n-times continuously differentiable functions $f \in \Gamma(U, V)$ (respectively, n-times continuously differentiable with uniformly bounded differentials). If the target space is clear, we write simply $C^n(U)$ and $C_b^n(U)$. For a differentiable $f \in \Gamma(\mathbb{R}^n, \mathbb{R}^n)$, we write the divergence of f as $\langle \nabla_x, f(x) \rangle = \operatorname{div} f(x) = \sum_{i \in [n]} \partial_i f_i(x)$.

Kernel Theory. Throughout, let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive definite kernel function, and let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d \times d}$ be a positive definite matrix-valued kernel function. Let $\mathcal{H}$ be the reproducing kernel Hilbert space (RKHS) defined by k, and let $\mathcal{H}^d$ be the respective RKHS defined by K. The reproducing property in $\mathcal{H}^d$ requires that for $f \in \mathcal{H}^d$ and $x \in \mathbb{R}^d$, we have

$$\langle f(x), v \rangle_{\mathbb{R}^d} = \langle f, K(\cdot, x)v \rangle_{\mathcal{H}^d} \quad \text{for all } v \in \mathbb{R}^d. \tag{2}$$

Specifically, we will be interested in the case when $K = kI_d$. Then, it holds that $f \in \mathcal{H}^d$ if and only if $f_i \in \mathcal{H}$ for all $i \in [d]$ (see Appendix A.2).

Characterization of Stationarity. The generator $\mathcal{L}$ of a stochastic process defined by (1) is a linear operator describing the infinitesimal expected change of functions $f \in C_b(\mathbb{R}^d)$ along the process. For a sufficiently regular SDE (1), $\mathcal{L}$ acts as a differential operator on $f \in C^2(\mathbb{R}^d)$ and is given by

$$\mathcal{L}f(x) = \langle b(x), \nabla f(x) \rangle + \frac{1}{2} \operatorname{tr}\big(a(x) \nabla \nabla f(x)\big), \tag{3}$$

where $a(x) = \sigma(x)\sigma(x)^T$. A probability density μ is a stationary solution of the SDE (1) if and only if it satisfies the martingale problem,

$$\mathbb{E}_{X \sim \mu}[\mathcal{L}f(X)] = 0 \quad \text{for all } f \in C^2(\mathbb{R}^d). \tag{4}$$

The class of μ-targeted diffusions, i.e., diffusion processes defined by (1) with stationary density μ, admits a full characterization: the drift decomposes as

$$b(x) = \frac{1}{2\mu(x)} \langle \nabla, \mu(x)\,(a(x) + c(x)) \rangle, \tag{5}$$

with a skew-symmetric stream coefficient c (Gorham et al., 2019). In this case, the generator can be written as

$$(\mathcal{L}f)(x) = \frac{1}{2\mu(x)} \langle \nabla, \mu(x)(a(x) + c(x)) \nabla f(x) \rangle. \tag{6}$$

A stationary diffusion process is called reversible if its law is invariant under time reversal. This holds if and only if detailed balance holds, i.e., $\mu^{-1}\langle\nabla, \mu c\rangle = 0$ (Pavliotis, 2014, Prop. 4.5.). This term is an additive part of the drift in (5), and we refer to it as the non-reversible part of the drift. We refer to the supplement Appendix A.4 for the derivation and technical conditions.

Having characterized stationarity in terms of the generator $\mathcal{L}$, we now turn to how this condition can be used as a learning criterion.

Identifiability. The parameters of a stationary diffusion cannot be uniquely identified from its stationary density μ alone. In particular, an SDE of the form $dX_t = sb(X_t)dt + \sqrt{s}\sigma(X_t)dB_t$ has generator $s\mathcal{L}$ and, if a stationary distribution exists, it coincides with that of the original (unscaled) SDE. Hence, at best, $\mathcal{L}$ is identifiable from μ up to a multiplicative constant, and moreover, only through the product $\sigma\sigma^T$ (Sokol and Hansen, 2014, Example 5.5). To resolve this ambiguity in our setting, we fix the speed parameter and model $\sigma\sigma^T$ directly, rather than σ itself. For further results on identifiability of SDE parameters we refer to the literature (Sokol and Hansen, 2014; Wang et al., 2023; Dettling et al., 2023).

Kernel Deviation from Stationarity (Lorch et al., 2024, KDS). The KDS is defined via the supremum in the martingale problem (4) over the unit ball $\mathcal{H}_{\leq 1}$ of a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, yielding

$$\sqrt{\mathrm{KDS}(\mathcal{L}, \mu; \mathcal{F})} = \sup_{f \in \mathcal{H}_{\leq 1}} \mathbb{E}_{X\sim\mu}[\mathcal{L}f(X)]. \tag{7}$$

This expression is meaningful when $\mathcal{H}$ is chosen such that $\mathrm{core}(\mathcal{A}) \subseteq \mathcal{H} \subseteq C^2(\mathbb{R}^d)$, ensuring that the operator $\mathcal{L}$ acts appropriately on elements of $\mathcal{H}$. The *core* is an inclusion-minimal set of functions that guarantees the equivalence between μ being a stationary solution of (1) and of (4). For details we refer to the supplement Appendix A.4.

The *Stein-type kernel deviation from stationarity* (SKDS), introduced as our proposed loss function in Section 3, builds on the same principles but replaces ∇f in (3) by a function g in some $\mathcal{H}^d$. This generalization is inspired by the theory of Stein discrepancies, in particular the diffusion kernel Stein discrepancy (DKSD) introduced by Barp et al. (2019).

Stein Discrepancy. A Stein discrepancy compares two distributions Q and P by measuring deviations that vanish under P. Concretely, given a Stein operator $\mathcal{T}_P$ such that $\mathbb{E}_{X\sim P}[(\mathcal{T}_P g)(X)] = 0$ for all g in a suitable function class $\mathcal{G}$, one defines

$$S(Q, P; \mathcal{G}; \mathcal{T}_P) := \sup_{g\in\mathcal{G}} |\mathbb{E}_{X\sim Q}[(\mathcal{T}_P g)(X)]|\,. \tag{8}$$

This yields an integral probability metric (IPM) tailored to P, often more tractable than generic IPMs (Gorham and Mackey, 2015; Anastasiou et al., 2023).

A particularly useful choice arises when P is the stationary distribution of an SDE with generator $\mathcal{L}$. In this case, $\mathcal{T}_P = \mathcal{L}$, and the Stein identity $\mathbb{E}_{X\sim P}[(\mathcal{T}_P g)(X)] = 0$ coincides with the martingale problem in (4). This generator-based perspective forms the foundation of kernel Stein discrepancies and their diffusion variants, which we build on next.

3 STEIN-TYPE KERNEL DEVIATION FROM STATIONARITY

Building on the generator approach, we obtain from (6) the (first-order) *diffusion Stein operator* (Gorham et al., 2019)

$$\mathcal{D}_\mu^m g := \frac{1}{\mu}\langle\nabla, \mu m g\rangle, \tag{9}$$

where $g \in \Gamma(\mathbb{R}^d, \mathbb{R}^d)$, $m \in \Gamma(\mathbb{R}^d, \mathbb{R}^{d\times d})$ and μ is a C^1-density of P. Together with a kernel K, this yields the *diffusion kernel Stein discrepancy* (Barp et al., 2019).

In our setting, we have the matrix $m = a + c$, with a and c as in Section 2. For $g \in C^1(\mathbb{R}^d, \mathbb{R}^d)$, we rewrite

$$\begin{aligned}\mathcal{D}_\mu^{a+c} g &= \frac{1}{\mu}\langle\nabla, \mu(a+c)g\rangle \\ &= \frac{1}{\mu}\left(\langle\langle\nabla, \mu(a+c)\rangle, g\rangle + \mathrm{tr}\left((a+c)\nabla g\right)\right) \\ &= \langle 2b, g\rangle + \mathrm{tr}\left(a\nabla g\right) + \mathrm{tr}\left(c\nabla g\right),\end{aligned} \tag{10}$$

using standard vector calculus (see Appendix A.1).

Note that the additional term $\mathrm{tr}(c\nabla g)$ sets (10) apart from the diffusion generator $\mathcal{L}f = \langle b, \nabla f\rangle + \frac{1}{2}\mathrm{tr}(a\nabla^2 f)$. Indeed, for any $f \in C^2(\mathbb{R}^d)$, one has $\mathcal{D}_\mu^{a+c}\nabla f = 2\mathcal{L}f$, since $\mathrm{tr}(c\nabla\nabla f) = 0$.[1]

In contrast to b and σ, the skew-symmetric component c is not directly accessible during learning—similar to how the stationary density of the diffusion specified by b and σ is unknown. While μ and c are both not easily computed, they exist uniquely. In particular, c is uniquely determined by b and σ up to $c\nabla\log\mu + \langle\nabla, c\rangle$. The density μ solves the stationary Fokker-Planck equation (Huang et al., 2015); and the skew-symmetric matrix c is then defined via (5). Dropping

[1] In fact, $\mathrm{tr}(CB) = 0$ for any symmetric matrix B and skew-symmetric matrix C, because $\mathrm{tr}(CB) = \mathrm{tr}((CB)^T) = -\mathrm{tr}(BC) = -\mathrm{tr}(CB)$.

$\mathrm{tr}(c\nabla g)$ from $\mathcal{D}_\mu^{a+c}g$, or equivalently replacing ∇f by g in $\mathcal{L}$, motivates the following operator.

Definition 1 (SKDS Operator)**.** Let $b \in C^0(\mathbb{R}^d, \mathbb{R}^d)$ and $\sigma \in C^0(\mathbb{R}^d, \mathbb{R}^{d\times d})$. For $g \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ define

$$\mathcal{S}(g)(x) := 2\langle b(x), g(x)\rangle + \mathrm{tr}\big(\sigma(x)\sigma(x)^\top \nabla_x g(x)\big)\,.$$

For a matrix-valued function $A \in C^1(\mathbb{R}^d, \mathbb{R}^{d\times d})$ we extend column-wise,

$$(\mathcal{S}(A)(x))_i := \mathcal{S}(A_i)(x),$$

where A_i is the i-th column of A. We write $\mathcal{S}_P$ (respectively, $\mathcal{S}_p$) if the operator is associated with an SDE inducing a stationary distribution P (respectively, stationary density p).

The above operator gives rise to the *Stein-type kernel deviation from stationarity* (SKDS).

Definition 2 (Stein-type KDS)**.** Let $(X_t)_{t\geq 0}$ be a diffusion with drift $b \in \Gamma(\mathbb{R}^d, \mathbb{R}^d)$ and diffusion term $\sigma \in \Gamma(\mathbb{R}^d, \mathbb{R}^{d\times d})$, and let $\mathcal{S}$ be the associated SKDS operator. For a scalar-valued kernel k with RKHS $\mathcal{H}$, let $\mathcal{H}^d$ be the RKHS associated with $K = kI_d$ with unit ball $\mathcal{H}^d_{\leq 1}$. For a distribution μ, the SKDS is defined as

$$\sqrt{\mathrm{SKDS}(\mathcal{S}, \mu; \mathcal{H}^d_{\leq 1})} := \sup_{g\in\mathcal{H}^d_{\leq 1}} \mathbb{E}_{X\sim\mu}[\mathcal{S}g(X)]. \tag{11}$$

Following the approach of Lorch et al. (2024, Example Section 3.3), we provide a simple illustration of how gradient descent on the Stein-type KDS can be used to infer the correct model parameters.

Example 3 (SKDS as a learning objective)**.** Consider the one-dimensional linear SDE $dX_t = b(x-\alpha)\,dt + \sigma\,dB_t$, which has stationary distribution $\mathcal{N}(\alpha, -\frac{\sigma^2}{2b})$ whenever $b < 0$, $\sigma > 0$ (Jacobsen, 1993). Suppose the target distribution is $\mathcal{N}(1, \frac{1}{2})$, corresponding to the stationary solution of the SDE $dX_t = -4(x-1)\,dt + 2\,dB_t$. For illustration, in Fig. 2, we compare to two alternatives: (red) $\alpha = \frac{1}{4}$, $b = -4$, $\sigma = 2$, and (blue) $\alpha = 1$, $b = -4$, $\sigma = 1$. Fig. 2 shows that SKDS correctly identifies the true parameters by attaining vanishing partial derivatives at the optimum.

3.1 Closed Form Representation

For sufficiently regular drift, diffusion, and kernel functions, the function in (11) can be expressed as an inner product in $\mathcal{H}^d$ with a representer function $g_{\mathcal{S},\mu} \in \mathcal{H}^d$. To state the fact, we introduce some notation. For a differentiable kernel $k(x,y)$, we write $\partial_{x_j}k(x,x)$ for the derivative with respect to the j-th entry of both arguments. We denote by $\mathcal{S}_i k(x,y)$, $i = \{1,2\}$, the SKDS operator applied to the i-th argument. Finally, $L^2_p(\mathbb{R}^m, \mathbb{R}^n)$ is the space of measurable functions $f : \mathbb{R}^m \to \mathbb{R}^n$ with $\|f\|^2_{L^2_p} := \int \|f(x)\|^2\,\mathrm{d}p(x) < \infty$, for a probability measure p on $\mathbb{R}^m$.

Lemma 4 (Representer function $g_{\mathcal{S},\mu}$)**.** *Let μ be a probability density. Let b and σ be μ-square-integrable, i.e., in L^2_μ. Let $k(x,x)$ be continuously differentiable with k and $\partial_{x_j}k(x,x)$ μ-integrable. Then there exists a unique representer function $g_{\mathcal{S},\mu} \in \mathcal{H}^d$ such that*

$$\mathbb{E}_{X\sim\mu}[\mathcal{S}g(X)] = \langle g, g_{\mathcal{S},\mu}\rangle_{\mathcal{H}^d} \tag{12}$$

with explicit form $g_{\mathcal{S},\mu}(\cdot) = \mathbb{E}_{X\sim\mu}[\mathcal{S}_1 K(X,\cdot)]$.

The proof of this and later results in this paper are in the supplement (Appendix B). With the representer function and assumptions on the regularity of the drift and diffusion functions, SDKS admits a closed form.

Lemma 5 (SKDS Closed Form)**.** *We adopt the assumptions of Lemma 4 and additionally assume bounded b, σ, k and $\partial_{x_i,y_j}k(x,y)$. Moreover, let k be twice continuously differentiable. Then,*

$$\mathrm{SKDS}(\mathcal{S}, \mu; \mathcal{H}^d_{\leq 1}) = \|g_{\mathcal{S},\mu}\|^2_{\mathcal{H}^d}\,, \tag{13}$$

where the norm takes the following closed expression

$$\|g_{\mathcal{S},\mu}\|^2_{\mathcal{H}^d} = \mathbb{E}_{X\sim\mu}\left[\mathbb{E}_{Y\sim\mu}\left[\mathcal{S}_1\mathcal{S}_2 K(X,Y)\right]\right]. \tag{14}$$

Note that bounded functions b and σ ensure the boundedness of the operator $\mathcal{S}$, which is a sufficient condition to interchange $\mathcal{S}$ and the integral to receive the RHS in (14).

For a matrix-valued kernel of the form $K = kI_d$, we can simplify (14) even further as

$$\begin{aligned}\mathcal{S}_1\mathcal{S}_2 K(x,y) = \;& 4\langle b(x), b(y)k(x,y)\rangle \\ & + 2\langle b(x), a(y)^T\nabla_y k(x,y)\rangle \\ & + 2\langle b(y), a(x)^T\nabla_x k(x,y)\rangle \\ & + \mathrm{tr}\left(a(x)a(y)^T\nabla_x\nabla_y k(x,y)\right).\end{aligned} \tag{15}$$

3.2 Properties

We define $C^{n,n}$ (respectively, $C^{n,n}_b$) to be the set of functions $k \in \Gamma(\mathbb{R}^n \times \mathbb{R}^n)$ with $(x,y) \mapsto \nabla^l_x\nabla^l_x k(x,y)$ continuous (respectively, continuous and uniformly bounded) for all $l \in [n]$. The Wasserstein-2 distance is defined as $d_{\mathcal{W}_2}(\mu, p) = \inf_{X\sim\mu, Z\sim P}\mathbb{E}[\|X - Z\|_2]$.

The following result is inspired by the diffusion kernel Stein discrepancy, which can be upper bounded by the Wasserstein-2 distance (Gorham and Mackey, 2017, Proposition 9). In the respective result on the SKDS, an additional term dependent on the stream-coefficient c appears, which is non-accessible during learning.

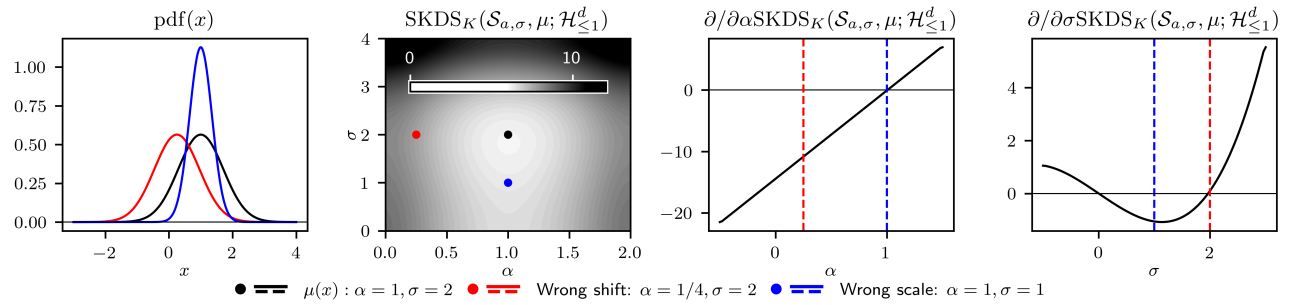

Figure 2: Stein-Type KDS for a stationary linear SDE: We applied SKDS to $n = 5000$ samples from the target distribution μ, using a one-dimensional Gaussian kernel with bandwidth 0.5. The models correspond to the SDE $dX_t = -4(X_t - \alpha)dt + \sigma dB_t$ for different values of α and σ. 1) PDFs of the stationary distributions. 2) An SKDS contour plot in α and σ. 3) and 4) Partial derivatives of the SKDS objective function for both model alternatives, showing that the gradient vanishes at the parameter values of the ground-truth model.

Proposition 6. *Let $\mathcal{S}_p$ be the SKDS operator associated to an SDE, $dX_t = b(X_t)dt + \sigma(X_t)dB_t$, with stationary density p supported on all of $\mathbb{R}^d$. Assume the following: (i) target probability density $\mu \in C^1$ supported on all of $\mathbb{R}^d$; (ii) $b \in C^1_b(\mathbb{R}^d, \mathbb{R}^d)$, and a, $c \in C^1_b(\mathbb{R}^d, \mathbb{R}^{d\times d})$; (iii) $K = kI_d$ with $k \in C^{(2,2)}_b(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$. Then*

$$\sqrt{\mathrm{SKDS}(\mathcal{S}_p, \mu; \mathcal{H}^d_{\leq 1})} \leq \lambda_k \lambda_{b,\sigma} \gamma(d_{\mathcal{W}_2}(\mu, p)) + \|k\|_\infty \|c\nabla \log p + \langle \nabla, c\rangle\|_{L^2_p(\mathbb{R}^d,\mathbb{R}^d)},$$

where $\gamma(x) = x + \sqrt{x}$, and λ_k, $\lambda_{b,\sigma}$ are constants depending on the regularity of k, b and σ.

The additional term $c\nabla \log p + \langle \nabla, c\rangle$ arises from the skew-symmetric stream coefficient c in the drift decomposition (5). It captures precisely the non-reversible part of the drift: when this term is zero, detailed balance holds and the diffusion is reversible (cf. (35)). Note that the stream coefficient c is not a free parameter in the last result. Under these assumptions it is uniquely determined by b and σ up to $c\nabla \log \mu + \langle \nabla, c\rangle$. Hence, c is a property of the SDE defined by b and σ.

The following result shows that the SKDS vanishes if and only if $\mathcal{S}$ is associated to a reversible diffusion with stationary density μ. In particular, vanishing SKDS implies that the stationary density of the corresponding diffusion matches the target density.

Theorem 7 (SKDS characterizes reversible diffusions)**.** *Let $\mathcal{S}$ be the SKDS operator associated to SDE $dX_t = b(X_t)dt + \sigma(X_t)dB_t$. Adopt the assumptions (i), (ii) and (iii) of Proposition 6. Moreover, assume (iv) $\sigma(x)\sigma(x)^T$ is positive definite for all $x \in \mathbb{R}^d$; (v) $C^\infty_c \subseteq \mathcal{H}$. Then $(X_t)_{t\geq 0}$ is a stationary reversible diffusion process with stationary density μ if and only if* $\mathrm{SKDS}(\mathcal{S}, \mu; \mathcal{H}^d_{\leq 1}) = 0$.

The assumptions on b and σ–namely (ii) and (iv)– ensure that $C^\infty_c(\mathbb{R}^d)$ forms a *core* of the generator $\mathcal{L}$ (Ethier and Kurtz, 1986). For the proof of Theorem 7, it is crucial that $\mathcal{H}^d$ contains this core. It is possible that more general core results could relax the required regularity assumptions on the drift and diffusion while still yielding Theorem 7. Furthermore, Lorch et al. (2024) argue in a related setting that, in practice, the conclusion extends informally to the case of merely continuous drift and diffusion functions. Continuous functions can become unbounded only as $\|x\| \to \infty$, whereas our studied process remains stationary.

Note that, up to this point, we have considered only a single target distribution. When applying SKDS to learn a causal framework, however, a parametrized model is trained jointly on multiple datasets – both observational and interventional (see Algorithm 1). In this joint setting, even a sufficiently expressive parametrized model need not admit a reversible solution. Despite a lack of theoretical guarantees characterizing the nature of a global solution in the case of multiple datasets, our approach as well as the approach by Lorch et al. (2024) perform well in experiments.

4 LEARNING STATIONARY DIFFUSIONS

4.1 Parametrization

Consider a parametrized SDE of form (1) with drift function b_θ and diffusion function σ_θ depending on parameters θ. Learning the parameters of a stationary diffusion in this setting means adjusting θ so that a stationary solution with distribution P exists and matches an observed target distribution μ. The asso-

ciated SKDS operator is denoted by $\mathcal{S}^\theta$. The following result and definition show the SKDS to be convex for a rich modeling class of drift and diffusion functions (for details see Appendix B).

Corollary 8 (Convexity of SKDS)**.** *Let $\mathcal{T} : C^1(\mathbb{R}^d, \mathbb{R}^d) \to C^0(\mathbb{R}^d, \mathbb{R}^l)$ be a linear operator independent of the parametrization $\theta \in \mathbb{R}^l$, $l \in \mathbb{N}$. Assume that the parametrized SKDS operator $\mathcal{S}^\theta$ takes the form*

$$\mathcal{S}^\theta \cdot = \langle \theta, \mathcal{T} \cdot \rangle. \tag{16}$$

Then $\mathrm{SKDS}(\mathcal{S}^\theta, \mu; \mathcal{H}^d_{\leq 1})$ *is convex in θ.*

The class of SDEs which admit a linear representation of the SKDS operator is notably larger than affine linear SDEs (which define, e.g., Ornstein-Uhlenbeck processes). We introduce a parameterized class of SDEs for which the SKDS operator becomes linear in the model parameters.

Example 9 (Linear SDE parametrization)**.** Let $\mathcal{J} = \{j_i : \mathbb{R}^d \to \mathbb{R}\}_{i\in[l]}$ be a basis of a finite-dimensional function space, e.g., a set of polynomial basis functions. Define the feature map $j(x) = (j_1(x), \dots, j_l(x))^T \in \mathbb{R}^l$, and let $B^\theta \in \mathbb{R}^{d\times l}$ be a parameter matrix. The drift is then modeled as

$$b(x) = B^\theta j(x).$$

To preserve linearity in the parameters, we model the *diffusion* via the symmetric positive semi-definite (PSD) matrix function $a(x) = \sigma(x)\sigma(x)^T$ directly (recall the identifiability remarks in Section 2). Let $\mathcal{V} = \{v_i\}_{i\in[m]} \subset \Gamma(\mathbb{R}^d, \mathbb{R}^d)$ be a set of vector-valued functions, and consider the set of rank-one PSD matrices $\{v_i(x)v_i(x)^T\}_{i\in[m]}$. Then, write $a(x)$ as a non-negative combination with parameters $A^\theta_i \geq 0$ as

$$a(x) = \sum_{i=1}^m A^\theta_i \, v_i(x)v_i(x)^T.$$

This ensures that $a(x)$ is PSD by construction. The SKDS operator takes then the form from (16) for

$$\mathcal{T}g(x) = \begin{pmatrix} \mathrm{vec}\left(j(x) \otimes g(x)\right) \\ V(x)\,\mathrm{vec}\left(\nabla g(x)\right) \end{pmatrix} \in \Gamma(\mathbb{R}^d, \mathbb{R}^{ld+m}), \tag{17}$$

where $\otimes$ denotes the outer product, and V stacks the vectorized basis diffusion matrices, i.e., $V(x) = (\mathrm{vec}(v_1(x)v_1(x)^T) \mid \dots \mid \mathrm{vec}(v_m(x)v_m(x)^T))^T$. For the calculations we refer to Appendix B.8.

The representational capacity of the above parameterized class is governed by the choices of basis functions in $\mathcal{J}$ and $\mathcal{V}$, which define the model's expressiveness. The dimensions l and m determine the number of parameters $(dl + m)$.

4.2 Stein-Type KDS as a Learning Objective

As the exact SKDS requires evaluation over the full support of μ, in applications, we turn to an empirical estimate from discrete samples. Let $D = \{x_1, \dots, x_N\}$ be an i.i.d. sample from the target distribution μ. An example for an unbiased empirical estimate of (14) is

$$\hat{\mathrm{SKDS}}(\mathcal{S}^\theta, D; \mathcal{H}^d_{\leq 1}) = \frac{1}{\lfloor N/2 \rfloor} \sum_{n=1}^{\lfloor N/2 \rfloor} \mathcal{S}^\theta_1 \mathcal{S}^\theta_2 K(x_{2n-1}, x_{2n}). \tag{18}$$

The computation of this estimator scales linearly with N. It is unbiased by linearity of expectation and the i.i.d. sampling. We specifically consider this linear statistic over the also possible U-statistic because its independent summands enable the simple bounds in Proposition 10. For benchmarking in Section 5, we also employ this linear estimator in the KDS-based approach proposed by Lorch.

By Corollary 8, the SKDS is convex in terms of a certain linear parametrization of the SDE. The following result shows that the empirical estimate (18) itself is 'almost convex' with high probability, depending on the regularity of the drift and diffusion function.

Proposition 10 (SKDS empirical estimate convex)**.** *We build on the setup from Example 9 and continue under the assumptions of Proposition 6. In addition, we assume that the basis functions are continuously differentiable and uniformly bounded, i.e., j, $v_i \in C^1_b$ with*

$$\|j(x)\|_\infty, \ \|v_i(x)v_i(x)^T\|_\infty \leq C \quad \forall i \in [m],$$

for some constant $C \in [1, \infty)$ and all $i \in [m]$. This ensures that the parametrized drift and diffusion satisfy the conditions of Proposition 6. Under these assumptions, the empirical estimate $\hat{\mathrm{SKDS}}(\mathcal{S}^\theta, D; \mathcal{H}^d_{\leq 1})$ *defined as in (18) is ε-strongly quasiconvex with high probability. In other words,*

$$\hat{\mathrm{SKDS}}(\mathcal{S}^\theta, D; \mathcal{H}^d_{\leq 1}) + \varepsilon \|\theta\|_2 \tag{19}$$

is convex in θ with probability greater than

$$1 - 2(dl + m) \exp\left(-\frac{\lfloor N/2 \rfloor \varepsilon^2/2}{L_2 + 2L_1\varepsilon/3}\right), \tag{20}$$

where L_1 and L_2 are constants depending on the bounds on j, v and the kernel k.

We remark that d enters the constants L_1 and L_2 with order $O(d^7)$ (see (76) and (77)). Consequently, for the lower bound in the previous proposition to remain meaningful, the number of samples N must scale on the order of $\Theta(d^8)$ in the dimensionality of the data.

Note that as $\theta \to 0$, the SKDS vanishes. This behavior is closely related to the speed scaling ambiguity in stochastic differential equations (see Section 2). To avoid this degeneracy and rule out the trivial minimizer $\theta = 0$, one can restrict the parameter space to the convex set $\{\theta \in \mathbb{R}^{dl+m} \mid \theta_1 = \alpha\}$ for some fixed constant α. The choice of α should be informed by domain knowledge: it must be non-zero and its sign must be appropriate for the problem at hand. This type of constraint is consistent with previous work, e.g., by fixing the mean reversion (Lorch et al., 2024, D.4).

4.3 Comparison with KDS

We have $\mathcal{L}u = \frac{1}{2}\mathcal{S}[\nabla u]$, which allows an analogous line of reasoning for both objectives. A comparison of their definitions in terms of operators, RKHSs, and closed-form expressions is given in Table 1. Notably, the KDS vanishes for Itô diffusions with stationary density μ, whereas the SKDS vanishes for those that are additionally reversible (Theorem 7).

We note that the generator $\mathcal{L}$ used in the KDS involves an extra differentiation step compared to $\mathcal{S}$, the operator underlying the SKDS. Hence, $\mathcal{L}_1\mathcal{L}_2 k(x, y)$ involves forth-order derivatives of k, whereas $\mathcal{S}_1\mathcal{S}_2 k(x, y)$ only needs derivatives of the kernel up to order two. For a direct comparison of the two general explicit forms we refer to (15) for the SKDS and to Appendix A.6 for the KDS. This explains why, as we demonstrate next, the SKDS is computationally cheaper than the KDS. To benchmark computation time, we randomly generate a linear drift b and linear diffusion σ, and mimic a gradient descent step by evaluating the KDS and SKDS with their respective kernels and computing derivatives with respect to b and σ. We consider three kernels: an RBF kernel (RBF), a tilted RBF kernel (Tilt. RBF), and an IMQ+ kernel (IMQ+); cf. Appendix C.1. Results show that the SKDS achieves a speedup over the KDS by a factor greater than 5 (Table 1).

5 SYNTHETIC EXPERIMENTS

A key objective of causal modeling is to predict the effects of interventions, including those not observed during training. To evaluate this capability, we assess how well learned models generalize to unseen interventions by comparing their predicted interventional distributions to ground truth. Models are trained on interventional data with known targets and tested on interventions affecting previously unaltered variables.

5.1 Setup

We follow the synthetic evaluation framework of Lorch et al. (2024); see their Section 6, Appendix, and codebase (Lorch, 2024) for details. The setup involves simulating dynamical systems, applying both training and test interventions, and evaluating predictive accuracy. We use the KDS-based causal model of Lorch et al. (2024) and replace it with our SKDS variant. Our implementation is available online (Bleile, 2025).

Data. We generate synthetic data from two system types: (i) sparse cyclic linear models (SCMs and stationary SDEs) and (ii) gene expression dynamics simulated with SERGIO (Dibaeinia and Sinha, 2020), a stationary nonlinear stochastic process over sparse acyclic networks, grounded in the chemical Langevin equation with Hill-type nonlinearities. Causal structures $G \in \{0, 1\}^{d\times d}$ are sampled from Erdős–Rényi (fixed edge probability, expected degree 3) and scale-free (preferential attachment, randomly directed) distributions. For each system, we create 50 datasets with 10 training and 10 test interventions, each targeting a distinct variable and containing 1000 samples. Interventions are additive shifts for both types of linear models and gene overexpression for SERGIO. During learning, we allow the more general shift–scale framework, modifying drift and diffusion (δ, β). All datasets are standardized by the observational mean and variance.

Models and interventions. The Stein-type KDS plugs naturally into the framework of (Lorch et al., 2024, Algorithm 1). We sketch the learning algorithm in Algorithm 1 and give the python code lines in Appendix D. The stationary diffusion processes are parameterized by either linear or multi-layer perceptron (MLP) mechanisms. The drift function $b_\theta(x)$ is defined component-wise for each variable x_j as follows:

$$\begin{aligned} \text{Linear model:} \quad & b_{\theta_j}(x)_j = b_j + w_j \cdot x \\ \text{MLP model:} \quad & b_{\theta_j}(x)_j = b_j + w_j \cdot \gamma(U_j x + v_j) - x_j \end{aligned}$$

where γ denotes the sigmoid activation function. The diffusion matrix is parameterized as a diagonal matrix $\sigma_\theta(x) = \mathrm{diag}(\exp(\sigma))$, with learned log-standard deviations. To remove the speed scaling invariance, we fix $w_j^j = -1$ in the linear and $u_j^j = 0$ in the MLP model (see Section 2 and the note after Proposition 10). Note that both model variants violate the boundedness assumptions on the drift and diffusion functions, introduced in Lemma 5 and subsequently used throughout all theoretical results. However, as discussed above, these assumptions can be informally relaxed in practice to merely require continuity.

Interventions ϕ are modeled as *shift-scale interven-*

	KDS	Stein-type KDS
Operator	$\mathcal{L}h = \langle b, \nabla_x h\rangle + \frac{1}{2}\operatorname{tr}\left(\sigma\sigma^T \nabla_x \nabla_x h\right); \quad h \in \Gamma(\mathbb{R}^d)$	$\mathcal{S}g = \langle b, g\rangle + \frac{1}{2}\operatorname{tr}\left(\sigma\sigma^T \nabla_x g\right); \quad g \in \Gamma(\mathbb{R}^d, \mathbb{R}^d)$
RKHS	$\mathbb{R}$-valued kernel k, RKHS $\mathcal{H} \subset \Gamma(\mathbb{R}^d)$	$\mathbb{R}^d$-valued kernel K, RKHS $\mathcal{H}^d \subset \Gamma(\mathbb{R}^d, \mathbb{R}^d)$
Definition	$\mathrm{KDS}(\mathcal{L}, \mu; \mathcal{H}_{\leq 1}) = \sup_{h \in \mathcal{H}_{\leq 1}} \mathbb{E}_{X\sim\mu}[\mathcal{L}h]$	$\mathrm{SKDS}(\mathcal{S}, \mu; \mathcal{H}^d_{\leq 1}) = \sup_{g \in \mathcal{H}^d_{\leq 1}} \mathbb{E}_{X\sim\mu}[\mathcal{S}g]$
Closed form	$\mathbb{E}_{X\sim\mu}\mathbb{E}_{Y\sim\mu}\left[\mathcal{L}_1\mathcal{L}_2 k(X, Y)\right]$	$\mathbb{E}_{X\sim\mu}\mathbb{E}_{Y\sim\mu}\left[\mathcal{S}_1\mathcal{S}_2 K(X, Y)\right]$
Time (ms)	(RBF) 117 ± 22 (Tilt.RBF) 183 ± 27 (IMQ+) 293 ± 37	(RBF) $\mathbf{22 \pm 5}$ (Tilt.RBF) $\mathbf{30 \pm 1}$ (IMQ+) $\mathbf{46 \pm 3}$

Table 1: Structural and computational comparison of KDS and Stein-type KDS on $n = 1000$ random samples in $d = 20$ dimensions, averaged over 50 repetitions.

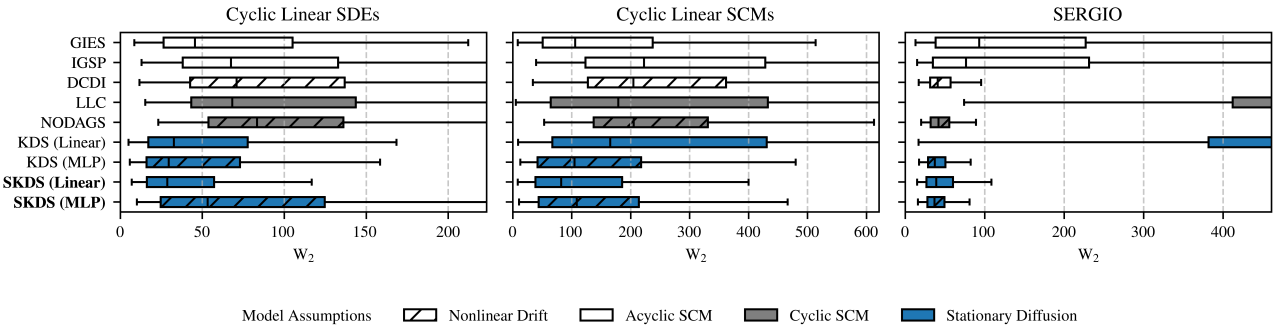

Figure 3: Benchmarking results for $d = 20$ variables with an Erdős–Rényi causal structure. The Wasserstein-2 distance ($\mathcal{W}_2$) was computed from 10 test interventions on unseen target variables across 50 randomly generated systems. Box plots depict the medians and interquartile ranges (IQR), with whiskers extending to the largest value within 1.5 times the IQR from the boxes.

tions on the known targets. For a variable x_j, the intervention given by $\phi = (\delta, \beta)$ modifies the drift and diffusion as

$$b_{\theta,\phi}(x)_j = b_\theta(x)_j + \delta, \qquad \sigma_{\theta,\phi}(x)_j = \beta \cdot \sigma_\theta(x)_j.$$

Algorithm 1: Learning Causal Stationary Diffusions via the Stein-type KDS

Input: Interventional datasets $\{D_1, \ldots, D_M\}$, kernel k, sparsity penalty λ, regularizer R
Output: Learned parameters θ and interventions $\{\phi_1, \ldots, \phi_M\}$
Initialize model θ and interventions $\{\phi_1, \ldots, \phi_M\}$;
while *not converged* **do**
  Sample environment index $i \sim \mathrm{Unif}([M])$;
  Sample batch $D \sim D_i$;
  Update θ and ϕ_i via gradient descent step $\propto -\nabla_{\theta,\phi_i}\left(\hat{\mathrm{SKDS}}(S^{\phi,\theta}, D; k) + \lambda R(\theta, \phi)\right)$;
return θ, $\{\phi_1, \ldots, \phi_M\}$

We benchmark SKDS against six established baselines. The graph discovery methods GIES (Hauser and Bühlmann, 2012) and IGSP (Wang et al., 2017) assume linear-Gaussian SCMs. They infer graph structure from interventional data, after which parameters are estimated in closed form. DCDI (Brouillard et al., 2020) extends this to nonlinear Gaussian SCMs parameterized by neural networks, jointly learning functional relations and noise variances. NODAGS (Sethuraman et al., 2023) relaxes acyclicity, modeling nonlinear cyclic SCMs via residual normalizing flows. LLC (Hyttinen et al., 2012) also addresses cyclic structures but in a linear-Gaussian setting, with sparsity-regularized optimization. Interventions are simulated in all SCM-based approaches by biasing the distribution of the target variables. Finally, we include the KDS-based stationary diffusion model (Lorch et al., 2024), which is based on the same SDE framework as SKDS and serves as the closest baseline. We refer to Lorch et al. (2024) for further implementation and tuning details.

Metrics. We evaluate the learned causal models via test interventions that shift the mean of the target

	SDE-ER		SCM-ER		SERGIO-ER	
Baseline	Lin	MLP	Lin	MLP	Lin	MLP
GIES	*		*		*	*
IGSP	*	*	*	*	*	*
DCDI	*	*	*	*		*
LLC	*	*	*	*	*	*
NODAGS	*	*	*	*	*	*
KDS (Linear)		*	*	*	*	*
KDS (MLP)		*	*			

Table 2: Significance tests for **SKDS (Linear)** (Lin) and **SKDS (MLP)** (MLP) on ER graphs with Wasserstein-2 distance. A black star indicates a statistically significant improvement of our method (margin $\geq 5\%$) over the considered baseline method, while a red star indicates the baseline significantly outperforming ours.

variables to match those observed in held-out interventional data. Performance is measured using the Wasserstein-2 distance ($\mathcal{W}_2$) between samples generated by the models under these interventions and the corresponding true interventional samples. This metric allows comparison across methods with either explicit or implicit density representations. Additionally, we report the mean squared error (MSE) between the predicted and true empirical means of the interventional distributions; compare (Lorch et al., 2024, D.2).

5.2 Results

In the main text, we report results for the Erdős–Rényi (ER) setting with Wasserstein distance as the evaluation metric; the full set of results and significance tests is provided in Appendix C.4. This focus is justified as the empirical behavior for ER and scale-free (SF) graphs is qualitatively similar. Moreover, in cyclic linear SDEs or SCMs all methods achieve comparably low MSE, while performance differences become apparent only once nonlinearity is introduced in the data. In the nonlinear case, MSE and Wasserstein distance lead to consistent conclusions.

The cyclic linear SDE setup aligns naturally with our stationary SDE formulation, and both SKDS variants perform competitively in this setting. In Fig. 3, we observe that the acyclic method GIES remains strong despite model misspecification, whereas cyclic approaches such as LLC and NODAGS underperform. SKDS (MLP) is outperformed by the other stationary diffusion models (see Table 2), which may reflect suboptimal hyperparameters or the inherent tradeoff between model complexity and optimization.

The results for the cyclic linear SCMs are largely similar to those for the cyclic linear SDEs, with stationary diffusion models and GIES performing best. This is particularly notable because the generated data satisfies the assumptions of LLC and NODAGS, suggesting that while these methods achieve reasonable MSE (see Fig. 4 in Appendix C.4) – possibly capturing mean behavior of interventional distributions – they struggle to accurately model full distributional properties, as reflected in the Wasserstein distance.

On the SERGIO gene expression data, which introduces realistic nonlinearities and model mismatch, methods with nonlinear drift dominate. Here, SKDS (MLP) is among the top performers, while SKDS (Linear) remains competitive despite its restricted functional form.

6 CONCLUSION

We introduced the Stein-type KDS as a principled framework for learning stationary diffusions. Theoretically, SKDS vanishes precisely for the class of stationary reversible diffusions whose stationary distribution matches the target. Furthermore, for linear parametrization (including expansions in nonlinear basis functions), the SKDS objective is convex with a quasiconvex empirical estimator. Empirically, SKDS matches or improves upon KDS and other competitive baselines across linear and nonlinear, cyclic and acyclic settings, while achieving a substantial reduction in computation time (Table 1). In summary, our findings highlight the robustness and flexibility of stationary diffusions as a computationally efficient foundation for causal modeling, providing a principled alternative to existing approaches.

Acknowledgements

The authors acknowledge support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 883818). Fabian Bleile acknowledges institutional support from the TUM Georg Nemetschek Institute for Artificial Intelligence for the Built World and the Munich Center for Machine Learning. Sarah Lumpp was further supported by the Munich Data Science Institute as well as the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Research, Technology and Space.

References

Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gret-

ton, Christophe Ley, Qiang Liu, et al. Stein's method meets computational statistics: A review of some recent developments. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, 38(1):120–139, 2023.

Ludwig Arnold. *Stochastic differential equations: theory and applications*. John Wiley & Sons, New York-London-Sydney, 1974.

Alessandro Barp, François-Xavier Briol, Andrew B. Duncan, Mark A. Girolami, and Lester W. Mackey. Minimum stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, volume 32, pages 12964–12976, 2019.

Fabian Bleile. Steinstadion: A framework for learning on diffusions. https://github.com/fbleile/steinstadion/tree/SKDS, 2025.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems*, volume 33, pages 21865–21877, 2020.

Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4(4):377–408, 2006.

Donald L. Cohn. *Measure theory*. Birkhäuser Advanced Texts: Basler Lehrbücher. Birkhäuser/Springer, New York, 2nd edition, 2013.

Philipp Dettling, Roser Homs, Carlos Améndola, Mathias Drton, and Niels Richard Hansen. Identifiability in continuous Lyapunov models. *SIAM Journal on Matrix Analysis and Applications*, 44(4): 1799–1821, 2023.

Payam Dibaeinia and Saurabh Sinha. Sergio: A single-cell expression simulator guided by gene regulatory networks. *Cell Systems*, 11(3):252–271.e11, 2020.

Stewart N. Ethier and Thomas G. Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986.

Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, volume 28, pages 226–234, 2015.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1292–1301, 2017.

Jackson Gorham, Andrew B Duncan, Sebastian J. Vollmer, and Lester Mackey. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928, 2019.

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

Wen Huang, Min Ji, Zhenxin Liu, and Yingfei Yi. Steady states of Fokker-Planck equations: I. Existence. *Journal of Dynamics and Differential Equations*, 27(3-4):721–742, 2015.

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

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

Heishiro Kanagawa, Alessandro Barp, Arthur Gretton, and Lester Mackey. Controlling moments with kernel stein discrepancies. *arXiv e-prints*, pages arXiv–2211, 2022.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 276–284, 2016.

Lars Lorch. Stadion: A framework for learning on diffusions. https://github.com/larslorch/stadion, 2024. Accessed: 2024-11-09.

Lars Lorch, Andreas Krause, and Bernhard Schölkopf. Causal Modeling with Stationary Diffusions. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pages 1927–1935. PMLR, 2024.

Bernt Øksendal. *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, 6th edition, 2003.

Grigorios A. Pavliotis. *Stochastic processes and applications*, volume 60 of *Texts in Applied Mathematics*. Springer, New York, 2014.

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, 2nd edition, 2009.

Stefano Pigola and Alberto G. Setti. *Global divergence theorems in nonlinear PDEs and geometry*, volume 26 of *Ensaios Matemáticos*. Sociedade Brasileira de Matemática, Rio de Janeiro, 2014.

Muralikrishnna G. Sethuraman, Romain Lopez, Rahul Mohan, Faramarz Fekri, Tommaso Biancalani, and Jan-Christian Huetter. Nodags-flow: Nonlinear cyclic causal structure learning. In *Proceedings of*

The 26th International Conference on Artificial Intelligence and Statistics, volume 206, pages 6371–6387. PMLR, 2023.

Alexander Sokol and Niels Richard Hansen. Causal interpretation of stochastic differential equations. *Electronic Journal of Probability*, 19:no. 100, 24, 2014.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2nd edition, 2000.

Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, 6:583–603, 1972.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

Joel A. Tropp. *An introduction to matrix concentration inequalities*, volume 8. Now Publishers, Inc., 2015.

Gherardo Varando and Niels Richard Hansen. Graphical continuous Lyapunov models. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020*, volume 124, pages 989–998. AUAI Press, 2020.

Yuanyuan Wang, Xi Geng, Wei Huang, Biwei Huang, and Mingming Gong. Generator identification for linear SDEs with additive and multiplicative noise. In *Advances in Neural Information Processing Systems*, volume 36, pages 64103–64138, 2023.

Yuhao Wang, Liam Solus, Karren D. Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems*, volume 30, pages 5822–5831, 2017.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   Yes, in section Sections 2 to 5.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   Important properties are stated in Sections 3 and 4. The complexity is highly dependent on the SDE model employed.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   Yes, see Bleile (2025). Moreover, refer to Lorch (2024) with the drop in replacement code given in Appendix D.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   Yes, assumptions are given at the beginning of each statement.
   (b) Complete proofs of all theoretical results.
   Yes, see Appendix B.
   (c) Clear explanations of any assumptions.
   Yes, we make clear in the proofs, where and why the assumptions are needed.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
   Yes, we refer to our code Bleile (2025) and the theoretical setup described in detail by Lorch et al. (2024); Lorch (2024).
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
   Yes, see above.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
   Yes, we sketch the metrics in Section 5 and refer for definitions to Lorch et al. (2024).
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
   Yes, see Appendix C.2.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets.
   Yes, we list the assets with their respective license under Appendix C.3.
   (b) The license information of the assets, if applicable.
   See above.
   (c) New assets either in the supplemental material or as a URL, if applicable.
   Yes, this is addressed in Appendix D.
   (d) Information about consent from data providers/curators.
   We use synthetic data, hence, not applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
   Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots.
   Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
   Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
   Not Applicable.

Supplementary Materials

A ADDITIONAL BACKGROUND

A.1 Vector Calculus

We give notation and important identities from vector calculus (based upon Barp et al. (2019, A.1)). For $f \in \Gamma(\mathbb{R}^n, \mathbb{R}^n)$, we denote $\langle \nabla_x, f(x) \rangle$ as the divergence of f, $\operatorname{div} f(x) = \sum_{i \in [n]} \partial_i f_i(x)$. For a function $g \in \Gamma(\mathcal{X})$, $v \in \Gamma(\mathcal{X}, \mathbb{R}^d)$ and $A \in \Gamma(\mathcal{X}, \mathbb{R}^{d \times d})$ with components A_{ij}, v_i, g, we have $(\nabla g)_i = \partial_i g$, $(\langle \nabla, A \rangle)_i = \langle \nabla, A_i \rangle$, where A_i is the i-th column of A. We define A^j as the j-th row of A respectively. Moreover, we have $(\nabla v)_{ij} = \partial_j v_i$, $\nabla^2 f \equiv \nabla(\nabla f)$. We have the following identities,

$$\begin{aligned}
\langle \nabla, gv \rangle &= \partial_i (g v_i) = v_i \partial_i g + g \partial_i v_i = \langle \nabla g, v \rangle + g \langle \nabla, v \rangle, \\
\langle \nabla, gA \rangle &= \partial_i (g A_{ij}) e_j = (A_{ij} \partial_i g + g \partial_i A_{ij}) e_j = \langle \nabla g, A \rangle + g \langle \nabla, A \rangle, \\
\langle \nabla, Av \rangle &= \partial_i (A_{ij} v_j) = \langle \langle \nabla, A \rangle, v \rangle + \operatorname{Tr}(A \nabla v),
\end{aligned}$$

where we treat $\langle \nabla, A \rangle$ and ∇g as column vectors. For a matrix $M \in \mathbb{R}^{m \times n}$, the column-wise vectorization of M is obtained by stacking its columns on top of each other, i.e.,

$$\operatorname{vec}(M) = \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_n \end{bmatrix} \in \mathbb{R}^{mn}.$$

A.2 Kernel Theory

This section reviews key concepts from kernel theory (Steinwart and Christmann, 2008; Carmeli et al., 2006).

Definition 11 (Positive Semidefinite Kernel)**.** Let X be an arbitrary set. A map $k : X \times X \to \mathbb{R}$ is called *positive semidefinite kernel* (PSD kernel) if and only if for all $n \in \mathbb{N}$ and all $x \in X^n$ the $n \times n$ matrix G with entries $G_{ij} := k(x_i, x_j)$ is positive semidefinite, i.e., $c^T G c \geq 0$ for all $c \in \mathbb{R}^n$.

Theorem 12 (PSD Kernels and Feature Maps)**.** *Let X be any set, and let $k : X \times X \to \mathbb{R}$.*

(i) Then k is a PSD kernel, if there is an inner product space $\mathcal{H}$ and a map $\phi : X \to \mathcal{H}$ such that

$$k(x, y) = \langle \phi(x), \phi(x) \rangle \quad \text{for all } x, y \in X. \tag{21}$$

(ii) Conversely, if k is a PSD kernel, then there exists a Hilbert space $\mathcal{H}$ and a map $\phi : X \to \mathcal{H}$ such that (21) holds.

For a given PSD kernel, the corresponding *feature map* ϕ and *feature space* $\mathcal{H}$ are not unique. However, there is a canonical choice for the feature space, a so-called reproducing kernel Hilbert space.

Definition 13 (Reproducing Kernel Hilbert Space)**.** Let X be a set, and let $\mathcal{H} \subseteq \mathbb{R}^X$ be an $\mathbb{R}$-Hilbert space[2] of functions on X with addition $(f + g)(x) := f(x) + g(x)$ and multiplication $(\lambda f)(x) := \lambda f(x)$. $\mathcal{H}$ is called a *reproducing kernel Hilbert space* (RKHS) on X if for all $x \in X$ the linear functional $\delta_x : \mathcal{H} \to \mathbb{R}$, $\delta_x(f) := f(x)$ is bounded (i.e., $\sup_{f \in \mathcal{H} \setminus \{0\}} \frac{|f(x)|}{\|f\|} < \infty$).

Note that since δ_x is linear, boundedness is equivalent to continuity. That is, the defining property of a RKHS is that evaluation of its functions at arbitrary points is continuous with respect to varying the function. Moreover, the RKHS $\mathcal{H}$ to a kernel k satisfies the *reproducing property* that $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$ for all $x \in \mathbb{R}^d$.

[2] i.e., inner-product space which is a complete metric space with its norm induced by the inner product.

Definition 14 (Universal Kernel)**.** A PSD kernel $k : X \times X \to \mathbb{R}$ on a metric space X is called *universal* if

$$E := \operatorname{span}\{k_x : X \to \mathbb{R} \mid k_x(y) = k(x, y),\, x \in X\}$$

is dense in the set of continuous functions with compact support, $C_c(X, \mathbb{R})$, i.e., for any $f \in C_c(X, \mathbb{R})$ and $\varepsilon > 0$, there exists $g \in E$ such that $\|g - f\|_\infty < \varepsilon$.

Note that if $\phi : X \to \mathcal{H}$ is a feature map corresponding to k, then for every $g \in E$ there exists $w \in \mathcal{H}$ such that $g = \langle w, \phi(\cdot)\rangle$.

This notion extends naturally to the vector-valued case. If $K(x, y) = k(x, y)I_d$ with k universal, then the associated RKHS is $\mathcal{H}^d := \mathcal{H} \times \cdots \times \mathcal{H}$, and finite linear combinations of kernel sections $K_x(\cdot)$ are dense in $C_c(X, \mathbb{R}^d)$. Hence K is universal in the vector-valued sense.

Definition 15 (Differentiability of a Kernel)**.** Let k be a kernel on $\mathbb{R}^n$ with RKHS $\mathcal{H}$. We adopt the notation from Steinwart and Christmann (2008) and regarding differentiability we interpret k as a function $\tilde{k} : \mathbb{R}^{2n} \to \mathbb{R}$. Then define $\partial_i \partial_{i+n} k = \partial_i \partial_{i+n} \tilde{k}$. More generally, define $\partial^{\alpha,\alpha} = \partial_1^{\alpha_1} \dots \partial_n^{\alpha_n} \partial_{n+1}^{\alpha_{n+1}} \dots \partial_{2n}^{\alpha_{2n}}$, where $\alpha \in \mathbb{N}^n$. We say k is m-times continuously differentiable, and write $k \in C^{(m,m)}$, if $\partial^{\alpha,\alpha} k$ exists and is continuous for all $\sum_{i\in[n]} \alpha_i \le m$.

The theory of kernels can be extended to the case where k maps to a Hilbert space $\mathcal{K}$ (Carmeli et al., 2006). The resulting RKHS consists of functions from X to $\mathcal{K}$. In particular, we are interested in the setting with $\mathbb{R}^d$-valued functions.

Definition 16 (RKHS for $\mathbb{R}^d$-valued functions)**.** (Carmeli et al., 2006, Def. 2.2) An $\mathbb{R}^d$-valued kernel of positive type on $X \times X$ is a map $K : X \times X \to \mathbb{R}^{d\times d}$ such that for all $N \in \mathbb{N}$, $x_1, \dots, x_N \in X$ and $c_1, \dots, c_N \in \mathbb{R}$, we have

$$\sum_{i,j\in[N]} c_i c_j \langle K(x_j, x_i)v, v\rangle_{\mathbb{R}^d} \ge 0 \quad \text{for all } v \in \mathbb{R}^d, \tag{22}$$

or equivalently, $K(x, y)$ is positive semi-definite for any x, $y \in X$.

Consider the pre-Hilbert space

$$\mathcal{H}_0^d = \operatorname{span}\left\{K(\cdot, x)v \mid x \in X, v \in \mathbb{R}^d\right\}, \tag{23}$$

with $f = \sum_i^n a_i K(\cdot, x_i)v_i \in \mathcal{H}_0^d$ for $a_i \in \mathbb{R}$, $x_i \in X$, and $v_i \in \mathbb{R}^d$ for all $i \in [n]$. A sesquilinear form on $\mathcal{H}_0^d \times \mathcal{H}_0^d$ is defined by

$$\langle f, g\rangle_{\mathcal{H}_0^d} = \sum_i^n \sum_j^m a_i b_j \langle K(y_j, x_i)v_i, w_j\rangle_{\mathbb{R}^d}; \tag{24}$$

compare Carmeli et al. (2006, Proposition 2.3). Then, the unique RKHS $\mathcal{H}^d$ with reproducing kernel K is defined as the completion of $\mathcal{H}_0^d$ with respect to $\langle\cdot,\cdot\rangle_{\mathcal{H}_0^d}$. This means

$$\begin{aligned}\mathcal{H}^d = \Big\{ f = \sum_i a_i K(\cdot, x_i)v_i \mid\ & a_i \in \mathbb{R},\ x_i \in X, \text{ and } v_i \in \mathbb{R}^d \text{ for all } i \in \mathbb{N} \text{ such that} \\ & \|f\|_{\mathcal{H}^d}^2 := \lim_{n\to\infty} \Big\| \sum_i^n a_i K(\cdot, x_i)v_i \Big\|_{\mathcal{H}_0^d}^2 = \sum_{i,j} a_i a_j \langle K(x_j, x_i)v_i, v_j\rangle_{\mathbb{R}^d} < \infty \Big\}.\end{aligned} \tag{25}$$

The reproducing property in an RKHS $\mathcal{H}^d$ defined by an $\mathbb{R}^d$-valued kernel K reads as follows. For $f \in \mathcal{H}^d$ and $x \in X$,

$$\langle f(x), v\rangle_{\mathbb{R}^d} = \langle f, K(\cdot, x)v\rangle_{\mathcal{H}^d}, \quad \text{for all } v \in \mathbb{R}^d. \tag{26}$$

We note that any $\mathbb{R}$-valued kernel k of positive type on $X \times X$ defines an $\mathbb{R}^d$-valued kernel of positive type on $X \times X$ together with a symmetric positive definite matrix B by $K = kB$. We denote the RKHS of k with $\mathcal{H}$.

Specifically, we will be interested in the case when $B = I_d$. Then, for $f \in \mathcal{H}^d$, it holds that $f_i \in \mathcal{H}$ as

$$\begin{aligned}\|f\|^2_{\mathcal{H}^d} &= \sum_{i,j} a_i a_j \langle K(x_j, x_i) v_i, v_j \rangle_{\mathbb{R}^d} \\ &= \sum_{i,j} a_i a_j k(x_j, x_i) \sum_{k \in [d]} (v_i)_k (v_j)_k \\ &= \sum_{k \in [d]} \sum_{i,j} a_i a_j k(x_j, x_i) (v_i)_k (v_j)_k \\ &= \sum_{k \in [d]} \|f_k\|^2_{\mathcal{H}} .\end{aligned}$$

A.3 Stochastic Differential Equations

Next, we provide a short introduction to stochastic differential equations; for further details we refer to Øksendal (2003).

Differential equations (DEs) are a natural choice to model deterministic processes and put (physical) quantities and their rates of change in relation. This approach may fail to accurately model a process if its randomness effects quantities too much. Discovered by the botanist Robert Brown, Brownian motion is the irregular thermal movement of small particles in liquids and gases. The physical quantities of such particles like speed and position were observed to be "too random" to be accurately modeled with DEs. A model that adjusts for randomness in a differential equation is called a Stochastic Differential Equation (SDE), and a common class of SDEs–Itô diffusions–extends ordinary differential equations by explicitly incorporating stochastic terms.

Definition 17 (Itô diffusion)**.** An Itô diffusion is a stochastic process $(X_t)_t$ that can be written in the form of

$$dX_t = b(X_t, t)dt + \sigma(X_t, t)dB_t, \quad t \in [0, T],\ X_0 = Z, \tag{27}$$

with $(B_t)_{t \geq 0}$ Brownian motion, $T > 0$, $b, \sigma : \mathbb{R}^n \times [0, T] \to \mathbb{R}^n$ measurable, and $Z \in L^2(\Omega)$ a square integrable random variable independent of the driving noise process. Define an autonomous (or time-homogeneous) Itô diffusion through

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t, \tag{28}$$

where the drift b and the diffusion (or noise) σ only depend on X_t and not on past values X_s for $s < t$. This ensures a strong solution $(X_t)_{t \geq 0}$ of (27), if it exists, to be Markov.

If $\sigma(x) = \sigma$ then the noise is called additive, otherwise multiplicative. In the modeling of physical systems using SDEs, additive noise is usually due to thermal fluctuations, whereas multiplicative noise is due to noise in some control parameter.

Example 18 (Ornstein-Uhlenbeck Process)**.** An Ornstein-Uhlenbeck process is defined by the SDE

$$dX_t = M(X_t - a)dt + DdB_t \tag{29}$$

with invertible drift matrix $M \in \mathbb{R}^{n \times n}$ and diffusion matrix $D \in \mathbb{R}^{n \times n}$. If the matrix M is *stable*, i.e., its eigenvalues have negative real parts, the affine linear drift introduces a restoring force towards the mean $a \in \mathbb{R}^n$. For stable M and positive definite DD^T, the Ornstein-Uhlenbeck process has a unique stationary solution $\mathcal{N}(a, \Sigma)$, where Σ solves $M\Sigma + \Sigma M^T = -DD^T$ (Arnold, 1974, Theorem (8.2.12)).

A.4 Generator and the Martingale Problem

The generator $\mathcal{A}$ of a stochastic process defined by (1) = (28) is a linear operator that acts on a function $f \in C_b(\mathbb{R}^n)$ describing the infinitesimal expected change when it is applied to the stochastic process. For $f \in C^2(\mathbb{R}^n)$ and sufficiently regular b and σ, the generator takes the form of a differential operator,

$$\mathcal{L}(f)(x) = \langle b(x), \nabla f(x) \rangle + \frac{1}{2} \operatorname{tr}\left(a(x) \nabla \nabla f(x)\right), \tag{30}$$

where $a(x) = \sigma(x)\sigma(x)^T$ is called the *covariance coefficient function*. The martingale problem (see Ethier and Kurtz (1986)) connects the generator $\mathcal{A}$ of an SDE (1) with its stationary solution P. In specific, by the arguments

in Ethier and Kurtz (1986, chapter 4), if (1) has a unique strong solution then μ is a stationary solution if and only if

$$\mathbb{E}_{X\sim\mu}[\mathcal{A}f(X)] = 0 \text{ for all } f \in \mathcal{D}. \tag{31}$$

Importantly, the domain $\mathcal{D}$ of $\mathcal{A}$ can be replaced by a smaller subset, called the *core* of $\mathcal{A}$, without affecting the result. For sufficiently regular drift b and diffusion σ such a core can be shown to be the space of smooth functions with compact support C_c^∞ (see Lorch et al. (2024, Appendix B.3)). As $C_c^\infty \subseteq C^2$, it follows that $\mathcal{A}$ takes the form of $\mathcal{L}$ in (31), which further simplifies the characterization.

The family of SDEs that admit a stationary solution with density μ are referred to as *μ-targeted diffusions*. Notably, the class of μ-targeted diffusions can be characterized: for continuously differentiable drift and diffusion functions, μ is a stationary density of a process solving (1) if and only if b takes the form $b(x) = \frac{1}{2\mu(x)}\langle\nabla, \mu(x)a(x)\rangle + h(x)$ for a non-reversible component $h \in C^1$ satisfying $\langle\nabla, \mu(x)h(x)\rangle = 0$ for all $x \in \mathbb{R}^d$. Moreover, if h is μ-integrable, then there exists a differentiable, μ-integrable, and skew-symmetric $d \times d$ matrix-valued function c, called *stream coefficient*, such that the differential operator $\mathcal{L}$ takes the form

$$(\mathcal{L}f)(x) = \frac{1}{2\mu(x)}\langle\nabla, \mu(x)(a(x) + c(x))\nabla f(x)\rangle \tag{32}$$

on $f \in C^2 \cap \mathrm{dom}(\mathcal{L})$. The drift then has the form

$$\begin{aligned} b(x) &= \frac{1}{2\mu(x)}\langle\nabla, \mu(x)(a(x) + c(x))\rangle \\ &= \frac{1}{2\mu(x)}\langle\nabla, \mu(x)a(x)\rangle + \underbrace{\frac{1}{2\mu(x)}\langle\nabla, \mu(x)c(x)\rangle}_{=h}. \end{aligned} \tag{33}$$

The non-reversible h can be related to the stationary probability flux $J = \mu h = \frac{1}{2}\langle\nabla, \mu(x)c(x)\rangle$, which – in stationarity – is the divergence-free part of the drift. The condition that the stationary probability flux vanishes, i.e.,

$$J = 0 \tag{34}$$

is called *detailed balance condition*. The detailed balance condition holds if and only if the diffusion is reversible (Pavliotis, 2014, Prop. 4.5.).

For later use, we can write h also in the form

$$h = \frac{1}{2}\left(c\nabla \log \mu + \langle\nabla, c\rangle\right). \tag{35}$$

We refer to Gorham et al. (2019); Pavliotis (2014) for the details.

A.5 Stein Discrepancy

Comparing probability distributions is essential for tasks like model evaluation, data generation, and hypothesis testing. A natural starting point is the maximum deviation between their expectations over a class of real-valued test functions $\mathcal{F}$ (Gorham and Mackey, 2015), given by

$$d_{\mathcal{F}}(Q, P) := \sup_{f\in\mathcal{F}} \left|\mathbb{E}_{X\sim Q}[f(X)] - \mathbb{E}_{X\sim P}[f(X)]\right|. \tag{36}$$

Stein discrepancies are a family of statistical divergences, defined as maximal deviations arising from Stein's method (Stein, 1972), that are particularly useful due to their computability. The following definition makes precise the notion of Stein discrepancy alluded to in Section 2.

Definition 19 (Stein Discrepancy (Gorham and Mackey, 2015))**.** Let P, Q be probability distributions on a set $\mathcal{X}$. A Stein discrepancy is defined via a real-valued operator $\mathcal{T}_P$, which acts on a set $\mathcal{G}$ of $\mathbb{R}^d$-valued functions with domain $\mathcal{X}$, such that we have a Stein identity:

$$\mathbb{E}_{X\sim P}[(\mathcal{T}_P g)(X)] = 0 \quad \text{for all } g \in \mathcal{G}. \tag{37}$$

The Stein discrepancy is then defined as

$$S(Q, P; \mathcal{G}; \mathcal{T}_P) := \sup_{g \in \mathcal{G}} |\mathbb{E}_{X\sim Q}[(\mathcal{T}_P g)(X)]| \,. \tag{38}$$

Stein discrepancies take the form as in (36), where $\mathcal{F}$ is designed to "zero out" any discrepancy under P, i.e., $\mathbb{E}_{X\sim P}[f(X)] = 0$ for all $f \in \mathcal{F}$. We see that by using (37) and calculate

$$\begin{aligned} &\sup_{g\in\mathcal{G}} |\mathbb{E}_{X\sim Q}[(\mathcal{T}_P g)(X)]| \\ &= \sup_{g\in\mathcal{G}} |\mathbb{E}_{X\sim Q}[(\mathcal{T}_P g)(X)] - \mathbb{E}_{X\sim P}[(\mathcal{T}_P g)(X)]| \\ &= d_{\mathcal{T}_P\mathcal{G}}(Q, P), \end{aligned} \tag{39}$$

where $\mathcal{T}_P\mathcal{G} := \{\mathcal{T}_P g \in \Gamma(\mathcal{X}) \mid g \in \mathcal{G}\}$. The operator $\mathcal{T}_P$ is called the *Stein operator*, while $\mathcal{G}$ is referred to as the *Stein set*. The Stein discrepancy is an integral probability metric (IPM) if and only if $\mathcal{T}_P\mathcal{G}$ forms a valid class of test functions for an IPM. Stein discrepancies are particularly attractive because they are often more computationally tractable than general integral probability metrics. One of their key advantages is flexibility: the choice of Stein operator $\mathcal{T}_P$ can be tailored to the specific problem at hand. For instance, in many applications, $\mathcal{T}_P$ can be defined in terms of the score function of P, which is especially useful when only an unnormalized density of P is available. Additionally, unlike some statistical divergences that require explicit density estimates, the Stein discrepancy depends on Q only through expectations, allowing it to be computed even when Q is represented by an empirical sample. Stein's method has led to significant advances in computational statistics in recent years. For a review of applications we refer to Anastasiou et al. (2023).

A.6 Kernel Deviation from Stationarity

The general explicit form of the KDS is given as

$$\begin{aligned} \mathcal{L}_1\mathcal{L}_2 k(x,y) = {} & b(x)\cdot\nabla_x\nabla_y k(x,y)\cdot b(y) \\ & + \tfrac{1}{2}\, b(x)\cdot\nabla_x \operatorname{tr}\big(a(y)\nabla_y\nabla_y k(x,y)\big) \\ & + \tfrac{1}{2}\, b(y)\cdot\nabla_y \operatorname{tr}\big(a(x)\nabla_x\nabla_x k(x,y)\big) \\ & + \tfrac{1}{4}\, \operatorname{tr}\Big(a(x)\,\nabla_x\nabla_x^\top\, \operatorname{tr}\big(a(y)\nabla_y\nabla_y k(x,y)\big)\Big). \end{aligned} \tag{40}$$

Note that fourth-order derivatives of the kernel are introduced. The general explicit form of the SKDS (15) only introduces derivatives of k up to order two.

B PROOFS

B.1 Proof of Lemma 4

Proof. We first show that if $\mathbb{E}_{X\sim\mu}[\mathcal{S}g(X)]$ is a continuous linear functional on $\mathcal{H}^d$, then, by the Riesz representation theorem, there exists a unique $g_{S,\mu} \in \mathcal{H}^d$ such that (12) holds for all $g \in \mathcal{H}^d$. An explicit form of $g_{\mathcal{S},\mu}$ is obtained by using the reproducing property,

$$\langle g_{\mathcal{S},\mu}(y), e_i\rangle = \langle g_{\mathcal{S},\mu}, K(\cdot,y)e_i\rangle_{\mathcal{H}^d} = \mathbb{E}_{X\sim\mu}[\mathcal{S}_1(K(X,y)e_i)], \tag{41}$$

where e_i is the i-th unit vector of $\mathbb{R}^d$. Hence, by the definition of the SKDS operator,

$$g_{\mathcal{S},\mu}(y) = \mathbb{E}_{X\sim\mu}[\mathcal{S}_1 K(X,y)]. \tag{42}$$

We verify that $\mathbb{E}_{X\sim\mu}[\mathcal{S}g(X)]$ is a continuous linear functional. As it is the concatenation of an expectation and $\mathcal{S}$, which are both linear, it remains to show that $\mathbb{E}_{X\sim\mu}[\mathcal{S}g(X)]$ is continuous on $\mathcal{H}^d$. This is the case if $\mathbb{E}_{X\sim\mu}[\mathcal{S}g(X)]$ is bounded on the unit ball of $\mathcal{H}^d$. By Jensen's inequality and Cauchy-Schwarz,

$$\begin{aligned} |\mathbb{E}_{X\sim\mu}[\mathcal{S}g(X)]| &= \left|\mathbb{E}_{X\sim\mu}[2\langle b(X), g(X)\rangle + \operatorname{tr}\left(\sigma(X)\sigma(X)^T\nabla_x g(X)\right)]\right| \\ &\le \mathbb{E}_{X\sim\mu}[2\,\|b(X)\|_2\,\|g(X)\|_2 + \left\|\sigma(X)\sigma(X)^T\right\|_F \|\nabla_x g(X)\|_F]. \end{aligned}$$

We leverage the reproducing property (26) and bound

$$\begin{aligned}\|g(x)\|_2^2 &= \sum_{i\in[d]} |\langle g, k(\cdot,x)e_i\rangle_{\mathcal{H}^d}|^2 \\ &\leq \|g\|_{\mathcal{H}^d}^2 \sum_{i\in[d]} \|k(\cdot,x)e_i\|_{\mathcal{H}^d}^2 = d\,\|g\|_{\mathcal{H}^d}^2\, k(x,x).\end{aligned} \tag{43}$$

Moreover, we use Steinwart and Christmann (2008, Cor. 4.36) to upper bound $\|\nabla_x g(X)\|_F$, as

$$\begin{aligned}\|\nabla_x g(x)\|_F^2 &= \sum_{i,j\in[d]} |\partial_j g_i(x)|^2 \\ &\leq \sum_{i,j\in[d]} \|g_i\|_{\mathcal{H}}^2\, \partial_{x_j} k(x,x) = d\,\|g\|_{\mathcal{H}^d}^2 \sum_{j\in[d]} \partial_{x_j} k(x,x).\end{aligned} \tag{44}$$

Then the boundedness of $\mathbb{E}_{X\sim\mu}[\mathcal{S}g(X)]$ on the unit ball of $\mathcal{H}^d$ follows from plugging in the bounds and using the assumptions on the integrands being square integrable with respect to μ,

$$|\mathbb{E}_{X\sim\mu}\left[\mathcal{S}g(X)\right]| \leq \sqrt{d}\,\|g\|_{\mathcal{H}^d}\,\mathbb{E}_{X\sim\mu}\left[2\,\|b(X)\|_2\,\sqrt{k(X,X)} + \left\|\sigma(X)\sigma(X)^T\right\|_F \sqrt{\sum_{j\in[d]}\partial_{x_j}k(x,x)}\right].$$

We verify that the expectation is indeed finite under the integrability assumptions. Using Cauchy-Schwarz inequality, we get

$$\mathbb{E}_{X\sim\mu}\left[2\,\|b(X)\|_2\,\sqrt{k(X,X)}\right] \leq 2\mathbb{E}_{X\sim\mu}\left[\|b(X)\|_2^2\right]\mathbb{E}_{X\sim\mu}\left[k(X,X)\right],$$

which is finite as $b\in L_\mu^2$ and $k\in L_\mu^2$. In the same way we calculate

$$\mathbb{E}_{X\sim\mu}\left[\left\|\sigma(X)\sigma(X)^T\right\|_F \sqrt{\sum_{j\in[d]}\partial_{x_j}k(x,x)}\right] \leq \mathbb{E}_{X\sim\mu}\left[\left\|\sigma(X)\sigma(X)^T\right\|_F^2\right]\mathbb{E}_{X\sim\mu}\left[\sum_{j\in[d]}\partial_{x_j}k(x,x)\right].$$

The term on the RHS is again finite, because $\sigma\in L_\mu^2$ and $\partial_{x_j}k(x,x)\in L_\mu^1$. □

B.2 Proof of Lemma 5

Proof. The supremum of $\mathbb{E}_{X\sim\mu}\left[\mathcal{S}g(X)\right]$ over the unit ball $\mathcal{H}^d_{\leq 1}$ can be expressed in terms of the representer function $g_{\mathcal{S},\mu}$ as

$$\sup_{g\in\mathcal{H}^d_{\leq 1}} \mathbb{E}_{X\sim\mu}\left[\mathcal{S}g(X)\right] = \sup_{g\in\mathcal{H}^d_{\leq 1}} \langle g, g_{\mathcal{S},\mu}\rangle_{\mathcal{H}^d} = \langle \frac{g_{\mathcal{S},\mu}}{\|g_{\mathcal{S},\mu}\|_{\mathcal{H}^d}}, g_{\mathcal{S},\mu}\rangle_{\mathcal{H}^d} = \|g_{\mathcal{S},\mu}\|_{\mathcal{H}^d}.$$

Using the representer property of $g_{\mathcal{S},\mu}$ from (12) and plugging in the explicit form of $g_{\mathcal{S},\mu}$ given in (42), we obtain

$$\begin{aligned}\|g_{\mathcal{S},\mu}\|_{\mathcal{H}^d}^2 &= \langle g_{\mathcal{S},\mu}, g_{\mathcal{S},\mu}\rangle_{\mathcal{H}^d} \\ &= \mathbb{E}_{X\sim\mu}\left[\mathcal{S}g_{\mathcal{S},\mu}(X)\right] \\ &= \mathbb{E}_{X\sim\mu}\left[\mathcal{S}_1\left[\mathbb{E}_{Y\sim\mu}\left[\mathcal{S}_2 K(X,Y)\right]\right]\right].\end{aligned}$$

Here, the subscript in $\mathcal{S}_1$ indicates that the operator is applied to the first and $\mathcal{S}_2$ to the second argument. We seek to interchange the expectation $\mathbb{E}_{Y\sim\mu}$ and application of $\mathcal{S}_1$. By Cohn (2013) and the μ-integrability of $\mathcal{S}_2K(X,Y)$, the integration and the application of $\mathcal{S}_1$ may be interchanged if $\mathcal{S}_1$ is a continuous linear operator on the unit ball B of $C_b^1(\mathbb{R}^d,\mathbb{R}^d)$ with respect to the norm $\|\cdot\|_b = \sup_{x\in\mathbb{R}^d}\|\cdot\|_2 + \sup_{x\in\mathbb{R}^d}\|\nabla_x\cdot\|_F$. Linearity is clearly satisfied, so we only need to show continuity.

Let $f\in C_b^1(\mathbb{R}^d,\mathbb{R}^d)$. The boundedness of $\mathcal{S}$ on B follows by the boundedness of b and σ as

$$|\mathcal{S}f| \leq 2\,\|b(x)\|_2\,\|f(x)\|_2 + \left\|\sigma(x)\sigma(x)^T\right\|_F\|\nabla_x f(x)\|_F \leq \|f\|_b\left(2\,\|b(x)\|_2 + \left\|\sigma(x)\sigma(x)^T\right\|_F\right).$$

Hence, we can write

$$\|g_{\mathcal{S},\mu}\|_{\mathcal{H}^d}^2 = \mathbb{E}_{X\sim\mu}\left[\mathbb{E}_{Y\sim\mu}\left[\mathcal{S}_1\mathcal{S}_2K(X,Y)\right]\right].$$

□

B.3 Proof of (15)

Proof. Let K_y denote the kernel function $K(\cdot, y)$ for a fixed y. By the definition of $\mathcal{S}$, we can write

$$\begin{aligned}(\mathcal{S}K_y(x))_i &= \langle 2b(x), K_y(x)_i\rangle + \operatorname{tr}\left(\sigma(x)\sigma(x)^T\nabla K_y(x)_i\right)\\ &= \langle 2b(x), e_i k(x,y)\rangle + \operatorname{tr}\left(\sigma(x)\sigma(x)^T\nabla(e_i k(x,y))\right))\\ &= 2b_i(x)k_y(x) + \operatorname{tr}\left(a(x)\left(e_i\nabla k_y(x)^T\right)\right)\\ &= 2b_i(x)k_y(x) + \operatorname{tr}\left(a_i(x)\nabla k_y(x)^T\right)\\ &= 2b_i(x)k_y(x) + \langle a_i(x), \nabla k_y(x)\rangle,\end{aligned}$$

where a_i is the i-th column of $a(x) = \sigma(x)\sigma(x)^T$. Thus, we can express the full operator as

$$\mathcal{S}K_y(x) = 2b(x)k_y(x) + a(x)^T\nabla k_y(x). \tag{45}$$

With this, $\mathcal{S}_1\mathcal{S}_2K(x,y)$ takes the closed form

$$\begin{aligned}\mathcal{S}_1\mathcal{S}_2K(x,y) &= \mathcal{S}_1\left[2b(y)k(x,y) + a(y)^T\nabla_y k(x,y)\right]\\ &= 4\langle b(x), b(y)k(x,y)\rangle\\ &\quad + 2\langle b(x), a(y)^T\nabla_y k(x,y)\rangle + 2\operatorname{tr}\left(a(x)b(y)\nabla_x^T k(x,y)\right)\\ &\quad + \operatorname{tr}\left(a(x)a(y)^T\nabla_x\nabla_y k(x,y)\right).\end{aligned} \tag{46}$$

By the trace rules we have

$$\operatorname{tr}\left(a(x)b(y)\nabla_x^T k(x,y)\right) = \operatorname{tr}\left(b(y)\nabla_x^T k(x,y)a(x)\right) = \langle b(y), a(x)^T\nabla_x k(x,y)\rangle,$$

and the statement follows. □

B.4 Proof of Proposition 6

Proof. In the first part of the proof, we verify the Stein identity for the diffusion Stein operator $\mathcal{D}$ from (9), i.e.,

$$\mathbb{E}_{Z\sim P}[\mathcal{D}_p^{a+c}g(Z)] = 0, \tag{47}$$

for all $g \in \mathcal{H}^d_{\leq 1}$ (Gorham and Mackey, 2017, Prop. 3). First, we establish that properties of the kernel K, or k, are inherited by functions in their respective RKH spaces. Specifically, by Steinwart and Christmann (2008, Cor. 4.23 and Cor. 4.36), $h \in \mathcal{H}$ is bounded and twice continuously differentiable with bounded derivatives, and hence, in specific, h has Lipschitz continuous gradients. These bounds depend only on k and $\|h\|_{\mathcal{H}}$ as

$$|h(x)| \leq \|h\|_{\mathcal{H}}\|k\|_\infty, \tag{48}$$

$$\left|\partial^i h(x)\right| \leq \|h\|_{\mathcal{H}}\sqrt{\partial^{i,i}k(x,x)}, \tag{49}$$

$$\left|\partial^j\partial^k h(x)\right| \leq \|h\|_{\mathcal{H}}\sqrt{\partial^{(j,k),(j,k)}k(x,x)}. \tag{50}$$

We define $\lambda_{\mathcal{H}} = \max_{i,j,k}\left(\|h\|_{\mathcal{H}}, \sqrt{\partial^{i,i}k(x,x)}, \sqrt{\partial^{(j,k),(j,k)}k(x,x)}\right)$. As $g \in \mathcal{H}^d$ if and only if $g_i \in \mathcal{H}$ for all $i \in [d]$, the preceding arguments apply to any $g \in \mathcal{H}^d$. We conclude that $\mathcal{H}^d_{\leq 1}$, the unit ball of $\mathcal{H}^d$, subsets the (scaled) classical Stein-Set

$$\mathcal{G}_{\|\cdot\|_2,\lambda_{\mathcal{H}}} = \left\{g : \mathbb{R}^d \to \mathbb{R}^d \mid \sup_{x,y\in\mathbb{R}^d, x\neq y}\max\left(\|g(x)\|_2, \|\nabla_x g(x)\|_2, \frac{\|\nabla_x g(x) - \nabla_x g(y)\|_2}{\|x-y\|_2}\right) \leq \lambda_{\mathcal{H}}\right\}. \tag{51}$$

Then, using Gorham and Mackey (2017, Prop. 3), it holds that $\mathbb{E}_{Z\sim P}[\mathcal{D}_p^{a+c}g(Z)] = 0$ for all $g \in \mathcal{H}^d_{\leq 1}$. The assumption that p is supported on all of $\mathbb{R}^d$ implicitly enters the proof in Gorham and Mackey (2017, Prop. 3) in an application of the divergence theorem when integrating over boundaries. [3]

[3] Alternatively to the full-support assumption, one could use the setting used in Barp et al. (2019, Proposition 1), where both p and μ are continuously differentiable, and the kernel K is assumed to be strictly integrally positive definite (IPD). The IPD condition ensures that $\mathbb{E}_{X\sim q, Z\sim q}[K(X,Z)] > 0$ for any non-zero measure q. In this setting one can weaken $\operatorname{supp}(p) = \mathbb{R}^d$ to $\mu > 0$ implies $p > 0$.

Fix a joint distribution $P_{X,Z}$ on (X, Z) such that the marginals are distributed as $X \sim \mu$ and $Z \sim p$. We then use the prior reasoning for a zero addition

$$E_{X\sim\mu}[\mathcal{S}_p g(X)] = \mathbb{E}_{(X,Z)\sim P_{X,Z}}[\mathcal{S}_p g(X) - \mathcal{D}_p^{a+c} g(Z)]. \tag{52}$$

By the definitions, and using the triangle inequality, Jensen's inequality and Cauchy-Schwarz, we get

$$\begin{aligned}
|E_{X\sim\mu}[\mathcal{S}_p g(X)]| &\leq \mathbb{E}\left[|2\langle b(X), g(X)\rangle + \operatorname{tr}(a(X)\nabla_x g(X)) - 2\langle b(Z), g(Z)\rangle - \operatorname{tr}(a(Z)\nabla_x g(Z))|\right] \\
&\quad + |\mathbb{E}[\operatorname{tr}(c(Z)\nabla_x g(Z))]| \\
&\leq \mathbb{E}[2\,|\langle b(X), g(X) - g(Z)\rangle| + 2\,|\langle b(X) - b(Z), g(Z)\rangle| \\
&\quad + \left|\langle \operatorname{vec}(a(X)^T), \operatorname{vec}(\nabla_x g(X) - \nabla_x g(Z))\rangle\right| + \left|\langle \operatorname{vec}(a(X)^T) - \operatorname{vec}(a(Z)^T), \operatorname{vec}(\nabla_x g(Z))\rangle\right|] \\
&\quad + |\mathbb{E}[\operatorname{tr}(c(Z)\nabla_x g(Z))]| \\
&\leq 2\mathbb{E}[\|b(X)\|_2 \|g(X) - g(Z)\|_2] + 2\mathbb{E}[\|b(X) - b(Z)\|_2 \|g(Z)\|_2] \\
&\quad + \mathbb{E}[\|\operatorname{vec}(a(X))\|_2 \|\operatorname{vec}(\nabla_x g(X) - \nabla_x g(Z))\|_2] + \mathbb{E}[\|\operatorname{vec}(a(X)) - \operatorname{vec}(a(Z))\|_2 \|\operatorname{vec}(\nabla_x g(Z))\|_2] \\
&\quad + |\mathbb{E}[\operatorname{tr}(c(Z)\nabla_x g(Z))]|\,,
\end{aligned} \tag{53}$$

where we write the trace as a vector product, i.e., $\operatorname{tr}(AB) = \langle \operatorname{vec}(A^T), \operatorname{vec}(B)\rangle$.

Building on the notation in Gorham and Mackey (2017), we define for any function $g \in \Gamma(\mathbb{R}^d, \mathbb{R}^d)$ the constants $M_0(g) = \sup_{x\in\mathbb{R}^d} \|g(x)\|_2$ and $M_1(g) = \sup_{x\neq y} \|g(x) - g(y)\|_2 / \|x - y\|_2$.

The individual terms in (53) can be bounded as follows. By Cauchy-Schwarz and the definition of $M_0(g)$ and $M_1(g)$, we obtain

$$\mathbb{E}[\|b(X)\|_2 \|g(X) - g(Z)\|_2] \leq \mathbb{E}[2M_0(b)M_0(g)]$$

and

$$\mathbb{E}[\|b(X)\|_2 \|g(X) - g(Z)\|_2] \leq \mathbb{E}[M_0(b)M_1(g) \|X - Z\|_2].$$

We combine these two upper bounds and use the fact that $\min(a, b) \leq \sqrt{ab}$ for $a,b \geq 0$. Then,

$$\begin{aligned}
\mathbb{E}[\|b(X)\|_2 \|g(X) - g(Z)\|_2] &\leq M_0(b)\mathbb{E}[\min(2M_0(g), M_1(g) \|X - Z\|_2)] \\
&\leq M_0(b)\mathbb{E}[\sqrt{2M_0(g)M_1(g) \|X - Z\|_2}] \\
&\leq M_0(b)\sqrt{2M_0(g)M_1(g)}\sqrt{\mathbb{E}[\|X - Z\|_2]}.
\end{aligned}$$

Similarly, we have

$$\mathbb{E}[\left\|\operatorname{vec}(a(X)^T)\right\|_2 \|\operatorname{vec}(\nabla_x g(X) - \nabla_x g(Z))\|_2] \leq M_0(\operatorname{vec}(a))\sqrt{2M_0(g)M_1(\operatorname{vec}(\nabla_x g))}\sqrt{\mathbb{E}[\|X - Z\|_2]},$$

as well as

$$\mathbb{E}[\|b(X) - b(Z)\|_2 \|g(Z)\|_2] \leq M_1(b)M_0(g)\mathbb{E}[\|X - Z\|_2],$$

and

$$\mathbb{E}[\|\operatorname{vec}(a(X)) - \operatorname{vec}(a(Z))\|_2 \|\operatorname{vec}(\nabla_x g(Z))\|_2] \leq M_1(\operatorname{vec}(a))M_0(\operatorname{vec}(\nabla_x g))\mathbb{E}[\|X - Z\|_2].$$

We denote

$$\begin{aligned}
\lambda_k &= \sup_{g\in\mathcal{H}^d_{\leq 1}} \max(M_0(g), M_1(g), M_0(\operatorname{vec}(\nabla_x g), M_1(\operatorname{vec}(\nabla_x g))), \text{ and} \\
\lambda_{b,\sigma} &= 2^{\frac{3}{2}} \max(M_0(b), M_1(b), M_0(\operatorname{vec}(a)), M_1(\operatorname{vec}(a))).
\end{aligned}$$

The finiteness of λ_k follows from the arguments in the proof of Gorham and Mackey (2017, Prop. 9). The finiteness of $\lambda_{b,\sigma}$ follows from the boundedness assumptions on the respective functions and their derivatives. As the inequalities hold for any $g \in \mathcal{H}^d_{\leq 1}$, we take the supremum over $\mathcal{H}^d_{\leq 1}$ on both sides to receive

$$\sqrt{\operatorname{SKDS}(\mathcal{S}, \mu; \mathcal{H}^d_{\leq 1})} \leq 2\lambda_k\lambda_{b,\sigma}(\sqrt{\mathbb{E}[\|X - Z\|_2]} + \mathbb{E}[\|X - Z\|_2]) + \sup_{g\in\mathcal{H}^d_{\leq 1}} \mathbb{E}[\operatorname{tr}(c(Z)\nabla_x g(Z))]. \tag{54}$$

The stated inequality now follows by taking the infimum of these bounds over all joint distributions (X, Z) with $X \sim \mu$ and $Z \sim p$, and by the definition of the Wasserstein-2 distance $d_{\mathcal{W}_2}(\mu, P) = \inf_{Z\sim P, X\sim\mu} \mathbb{E}[\|X - Z\|_2]$. The supremum involving the stream-coefficient c is upper bounded in Lemma 21. □

The second term on the right-hand side of (54) admits a closed form similarly to the SKDS due to the linearity in g. This can be seen by following the steps of Lemma 4, Lemma 5 and (15). In specific, Riesz' representation theorem is applicable due to the boundedness of k and c. However, as shown in the following lemma, under the assumption of a sufficiently rich RKHS, there is another, more interpretable, closed form of $\sup_{g\in\mathcal{H}^d_{\leq 1}} \mathbb{E}[\mathrm{tr}(c(Z)\nabla_x g(Z))]$.

The following Lemma combines several properties of RKHS from Steinwart and Christmann (2008, Chapter 4.3).

Lemma 20. *Let $\mathcal{H}^d$ be the RKHS with kernel $K = kI_d$ for an universal kernel $k \in C_b(\mathbb{R}^d \times \mathbb{R}^d)$. Moreover, let p be a probability density on $\mathbb{R}^d$ with respect to the Lebesgue measure. Then*

1. *$\mathcal{H}^d$ is dense in $L^2_p(\mathbb{R}^d, \mathbb{R}^d)$ with $\mathcal{H}^d \subseteq L^2_p(\mathbb{R}^d, \mathbb{R}^d)$.*
2. *The inclusion operator $\mathrm{id} : \mathcal{H}^d \to L^2_p(\mathbb{R}^d, \mathbb{R}^d)$, $\mathrm{id}(g) = g$, is bounded, i.e., $\|g\|_{L^2_p(\mathbb{R}^d,\mathbb{R}^d)} \leq C \|g\|_{\mathcal{H}^d}$ for all $g \in \mathcal{H}^d$.*

   The adjoint operator of the inclusion $\mathrm{id} : \mathcal{H}^d \to L^2_p(\mathbb{R}^d, \mathbb{R}^d)$ is defined by

$$S_k^\star : L^2_p(\mathbb{R}^d, \mathbb{R}^d) \to \mathcal{H}; \quad (S_k^\star g)(x) := \int_X K(x, x')\, g(x')\, p(x')\, dx'.$$

   From the boundedness and linearity of id it follows that $S_k^\star$ is linear and bounded as well.
3. *Fix $f \in L^2_p(\mathbb{R}^d, \mathbb{R}^d)$. Then*

$$\sup_{g\in\mathcal{H}^d_{\leq 1}} \langle g, f\rangle_{L^2_p(\mathbb{R}^d,\mathbb{R}^d)} = 0 \iff f \equiv 0 \quad p-a.e. \tag{55}$$

   In specific, $S_k^\star$ is injective.

Proof. 1. By assumption, k is universal, which implies that K is universal (see Definition 14 and the comment below). Hence, $\mathcal{H}^d$ is dense in $C_c(\mathbb{R}^d)$. As p is a finite Borel measure on $\mathbb{R}^d$, it holds that $C_c(\mathbb{R}^d, \mathbb{R}^d)$ is dense in $L^2_p(\mathbb{R}^d, \mathbb{R}^d)$ with respect to $\|\cdot\|_{L^2_p(\mathbb{R}^d,\mathbb{R}^d)}$. This gives the result. By assumption k is bounded, then Steinwart and Christmann (2008, Theorem 4.23) gives that any $g \in \mathcal{H}^d$ is bounded with $|g(x)| \leq \|g\|_{\mathcal{H}^d} \|k\|_\infty$. Hence, the inclusion operator $\mathrm{id} : \mathcal{H}^d \to L^2_p(\mathbb{R}^d, \mathbb{R}^d)$, $\mathrm{id}(g) = g$ is well-defined.

2. By the previous arguments, it follows $\|g\|_{L^2_p(\mathbb{R}^d,\mathbb{R}^d)} \leq \|g\|_{\mathcal{H}^d} \|k\|_\infty$, and hence id is bounded.

3. "$\Rightarrow$" For a fixed $f \in L^2_p(\mathbb{R}^d, \mathbb{R}^d)$ and using the adjoint, we can write

$$\sup_{g\in\mathcal{H}^d_{\leq 1}} \langle g, f\rangle_{L^2_p(\mathbb{R}^d,\mathbb{R}^d)} = \sup_{g\in\mathcal{H}^d_{\leq 1}} \langle g, S_k^\star f\rangle_{\mathcal{H}^d} = \|S_k^\star f\|_{\mathcal{H}},$$

   where the last step follows by the Cauchy–Schwarz inequality in $\mathcal{H}^d$. So the supremum is exactly the RKHS norm of $S_k^\star f$.

   Now use the relationship between the kernel of the adjoint and the range of the forward operator: for any bounded operator between Hilbert spaces we have $\ker(S_k^\star) = \big(\mathrm{range}\,\mathrm{id}\big)^\perp$. Because $\mathcal{H}^d$ is dense in $L^2_p(\mathbb{R}^d, \mathbb{R}^d)$, $\mathrm{range}\,\mathrm{id}$ is dense in $L^2_p(\mathbb{R}^d, \mathbb{R}^d)$, hence $\big(\mathrm{range}\,\mathrm{id}\big)^\perp = \{0\}$. Therefore $\ker(S_k^\star) = \{0\}$, so $S_k^\star f = 0$ implies $f = 0$ p-a.e. Moreover, $\|S_k^\star f\|_{\mathcal{H}} = 0$ implies $S_k^\star f = 0$ and the statement follows.

   "$\Leftarrow$" If $f = 0$ p-a.e. then trivially $\langle g, f\rangle_{L^2} = 0$ for all g, so the supremum is 0.

□

Lemma 21. *Let P be a probability distribution with continuously differentiable density p. Let $Z \sim P$, and let $c \in C^1_b(\mathbb{R}^d, \mathbb{R}^d \times \mathbb{R}^d)$ be P-integrable. Let $K = kI_d$ with $k \in C_b^{(1,1)}(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$. Then*

$$0 \leq \sup_{g\in\mathcal{H}^d_{\leq 1}} \mathbb{E}[\mathrm{tr}(c(Z)\nabla_x g(Z))] \leq \|k\|_\infty \|c\nabla \log p + \langle \nabla, c\rangle\|_{L^2_p(\mathbb{R}^d,\mathbb{R}^d)}.$$

If k is universal, then

$$\sup_{g \in \mathcal{H}^d_{\leq 1}} \mathbb{E}[\mathrm{tr}(c(Z)\nabla_x g(Z))] = \|S_k^\star(c\nabla \log p + \langle \nabla, c\rangle)\|_{\mathcal{H}^d}\,, \tag{56}$$

with

$$\sup_{g \in \mathcal{H}^d_{\leq 1}} \mathbb{E}[\mathrm{tr}(c(Z)\nabla_x g(Z))] = 0 \iff c\nabla \log p + \langle \nabla, c\rangle = p^{-1}\langle \nabla, pc\rangle = 0. \tag{57}$$

Proof. Let $Z \sim P$ with density p and let $g \in \mathcal{H}^d_{\leq 1}$. We apply integration by parts (IbP) on the term $\mathbb{E}[\mathrm{tr}(c(Z)\nabla_x g(Z))]$. To verify that IbP can be applied, it suffices to show that $gc \in L^1_p$ and $\langle \nabla, gc\rangle \in L^1_p$ (Pigola and Setti, 2014). This becomes clear by the following arguments. For $g \in \mathcal{H}^d_{\leq 1}$, it follows that $g \in C^1_b$ by the boundedness and differentiability of k (as before, e.g., (43) and (44)). Moreover it holds $c \in C^1_b$ by assumption.

We use the notation defined in Appendix A.1 when we denote the j-th row of c by c^j and the j-th column of c by c_j. Calculate

$$\begin{aligned}
\mathbb{E}_{Z\sim P}[\mathrm{tr}(c(Z)\nabla_x g(Z))] &= \sum_j \int_{\mathbb{R}^d} c^j(z)\nabla g_j(z) p(z) dz \\
&= -\sum_j \int_{\mathbb{R}^d} g_j(z)\langle \nabla, p(z)(c^j(z))^T\rangle dz && \text{(IbP)} \\
&= \sum_j \int_{\mathbb{R}^d} g_j(z)\langle \nabla, p(z)c_j(z)\rangle dz \\
&= \sum_j \int_{\mathbb{R}^d} g_j \left(\langle \nabla p, c_j\rangle + p\langle \nabla, c_j\rangle\right) dz && \text{(Vector Calculus Appendix A.1)} \\
&= \sum_j \int_{\mathbb{R}^d} g_j \left(\langle \nabla \log p, c_j\rangle + \langle \nabla, c_j\rangle\right) p dz \\
&= \langle g, c\nabla \log p + \langle \nabla, c\rangle\rangle_{L^2_p(\mathbb{R}^d,\mathbb{R}^d)}.
\end{aligned} \tag{58}$$

As $g \in \mathcal{H}^d_{\leq 1}$ is bounded, we have $g \in L^2_p(\mathbb{R}^d, \mathbb{R}^d)$. Hence,

$$\begin{aligned}
\sup_{g \in \mathcal{H}^d_{\leq 1}} \mathbb{E}[\mathrm{tr}(c(Z)\nabla_x g(Z))] &= \sup_{g \in \mathcal{H}^d_{\leq 1}} \langle g, c\nabla \log p + \langle \nabla, c\rangle\rangle_{L^2_p(\mathbb{R}^d,\mathbb{R}^d)} \\
&\leq \sup_{g \in \mathcal{H}^d_{\leq 1}} \|g\|_{L^2_p(\mathbb{R}^d,\mathbb{R}^d)} \|c\nabla \log p + \langle \nabla, c\rangle\|_{L^2_p(\mathbb{R}^d,\mathbb{R}^d)} \\
&\leq \sup_{g \in \mathcal{H}^d_{\leq 1}} \|g\|_{\mathcal{H}^d} \|k\|_\infty \|c\nabla \log p + \langle \nabla, c\rangle\|_{L^2_p(\mathbb{R}^d,\mathbb{R}^d)} \\
&= \|k\|_\infty \|c\nabla \log p + \langle \nabla, c\rangle\|_{L^2_p(\mathbb{R}^d,\mathbb{R}^d)}\,.
\end{aligned} \tag{59}$$

If k is universal, then, using the adjoint as in Lemma 20, we calculate

$$\begin{aligned}
\sup_{g \in \mathcal{H}^d_{\leq 1}} \mathbb{E}[\mathrm{tr}(c(Z)\nabla_x g(Z))] &= \sup_{g \in \mathcal{H}^d_{\leq 1}} \langle \mathrm{id}(g), c\nabla \log p + \langle \nabla, c\rangle\rangle_{L^2_p(\mathbb{R}^d,\mathbb{R}^d)} \\
&= \sup_{g \in \mathcal{H}^d_{\leq 1}} \langle \mathrm{id}(g), c\nabla \log p + \langle \nabla, c\rangle\rangle_{L^2_p(\mathbb{R}^d,\mathbb{R}^d)} \\
&= \sup_{g \in \mathcal{H}^d_{\leq 1}} \langle g, S_k^\star(c\nabla \log p + \langle \nabla, c\rangle)\rangle_{\mathcal{H}^d} \\
&= \|S_k^\star(c\nabla \log p + \langle \nabla, c\rangle)\|_{\mathcal{H}^d}\,.
\end{aligned} \tag{60}$$

An application of the previously obtained result Lemma 20 concludes the proof. □

Interestingly, the term $c\nabla \log p + \langle \nabla, c\rangle$ is (up to a constant factor) the non-reversible component mentioned in Appendix A.4, and, when multiplied with p, resolves to the stationary probability flux (Pavliotis, 2014).

B.5 Proof of Theorem 7

Proof. Before we start with the proof, note that assumption (v), $C_c^\infty \subseteq \mathcal{H}$ implies the universality of k, because C_c^∞ is dense in C_c.

"$\Rightarrow$": We assume that μ is the stationary solution to the SDE. As the diffusion process is reversible, the detailed balance condition holds (see (34)). In specific, it holds

$$\begin{aligned} 0 &= \frac{1}{\mu}\langle \nabla, \mu c\rangle \\ &= c\nabla \log \mu + \langle \nabla, c\rangle. \quad \text{(Vector Calculus Appendix A.1)} \end{aligned} \tag{61}$$

An application of Proposition 6 with $d_{\mathcal{W}_2}(\mu,\mu) = 0$ and $\|c\nabla \log p + \langle \nabla, c\rangle\|_{L^2_\mu(\mathbb{R}^d,\mathbb{R}^d)} = 0$, concludes the proof.

"$\Leftarrow$": We first show that a vanishing SKDS implies that μ is the stationary density of the diffusion process $(X_t)_{t\geq 0}$. We therefore follow the arguments showing that the KDS characterizes stationary diffusions with their stationary density matching the target density (Lorch et al., 2024, Theorem 3). Note that (ii) and (iv) implies that $C_c^\infty(\mathbb{R}^d)$ is a core of the generator $\mathcal{A}$ of the stochastic process $(X_t)_{t\geq 0}$ (Ethier and Kurtz, 1986, Theorem 1.6, p.370). The generator takes the form $\mathcal{L}\cdot = \langle b, \nabla \cdot\rangle + \frac{1}{2}\operatorname{tr}\left(\sigma\sigma^T \nabla\nabla \cdot\right)$ on $C^2(\mathbb{R}^d)$, and in specific on $C_c^\infty(\mathbb{R}^d)$. Hence a stationary solution μ is characterized by

$$\sup_{u\in C_c^\infty(\mathbb{R}^d)} \mathbb{E}_{X\sim\mu}[\mathcal{L}u] = 0. \tag{62}$$

Now, we have that $\mathcal{L}u = \frac{1}{2}\mathcal{S}[\nabla u]$. From $\operatorname{SKDS}(\mathcal{S},\mu;\mathcal{H}^d_{\leq 1}) = 0$ it follows that $\mathbb{E}_{X\sim\mu}[\mathcal{S}g] = 0$ for all $g \in \mathcal{H}^d_{\leq 1}$ since the supremum is non-negative. This expands to any $g \in \mathcal{H}$, as

$$\mathbb{E}_{X\sim\mu}[\mathcal{S}g] = \|g\|_{\mathcal{H}^d} \left\langle g/\|g\|_{\mathcal{H}^d}, g_{\mathcal{S},\mu}\right\rangle = \|g\|_{\mathcal{H}^d}\cdot 0 = 0,$$

where we used that $g/\|g\|_{\mathcal{H}^d} \in \mathcal{H}^d_{\leq 1}$. By the assumptions on the kernel, (iii) and (v), we have that $C_c^\infty(\mathbb{R}^d,\mathbb{R}^d) \subseteq \mathcal{H}^d$. Equation (62) then follows from the observation that $\left\{\nabla_x u \mid u \in C_c^\infty(\mathbb{R}^d)\right\} \subseteq C_c^\infty(\mathbb{R}^d,\mathbb{R}^d)$.

It remains to show that $(X_t)_{t\geq 0}$ is a *reversible* stationary diffusion. We use the Stein identity of the diffusion Stein operator $\mathcal{D}_\mu^{a+c}$ in the same way as before in Appendix B.4, i.e., $\mathbb{E}_{Z\sim\mu}[\mathcal{D}_\mu^{a+c} g(Z)] = 0$ for all $g \in \mathcal{H}^d_{\leq 1}$. Then, we calculate

$$\begin{aligned} E_{X\sim\mu}[\mathcal{S}_\mu g(X)] &= \mathbb{E}_{X\sim\mu}[\mathcal{S}_\mu g(X)] - \mathbb{E}_{Z\sim\mu}[\mathcal{D}_\mu^{a+c} g(Z)] \\ &= \mathbb{E}_X\left[2\langle b(X), g(X)\rangle + \operatorname{tr}(a(X)\nabla_x g(X))\right] \\ &\quad - \mathbb{E}_Z\left[2\langle b(Z), g(Z)\rangle + \operatorname{tr}(a(Z)\nabla_x g(Z)) + \operatorname{tr}(c(Z)\nabla_x g(Z))\right] \\ &= -\mathbb{E}_Z\left[\operatorname{tr}(c(Z)\nabla_x g(Z))\right]. \end{aligned} \tag{63}$$

This sets us up for

$$\begin{aligned} 0 &= \sqrt{\operatorname{SKDS}(\mathcal{S},\mu;\mathcal{H}^d_{\leq 1})} \\ &= \sup_{g\in\mathcal{H}^d_{\leq 1}} E_{X\sim\mu}[\mathcal{S}_\mu g(X)] \\ &= \sup_{g\in\mathcal{H}^d_{\leq 1}} \mathbb{E}_Z\left[\operatorname{tr}(c(Z)\nabla_x g(Z))\right] \quad (63) \\ &= \|S_k^\star(c\nabla \log p + \langle \nabla, c\rangle)\|_{\mathcal{H}^d}. \quad (56) \end{aligned} \tag{64}$$

We conclude $\mu^{-1}\langle \nabla, \mu c\rangle = 0$ using Lemma 21 (specifically (57)), and hence $(X_t)_{t\geq 0}$ is a reversible diffusion process with stationary density μ. □

B.6 Proof of Corollary 8

Proof. Assume that we have a linear parametrization $\mathcal{S}^\theta \cdot = \langle \theta, \mathcal{T}\cdot\rangle$. We aim to show that

$$\operatorname{SKDS}(\mathcal{S}^\theta,\mu;\mathcal{H}^d_{\leq 1}) = \left(\sup_{g\in\mathcal{H}^d_{\leq 1}} \mathbb{E}_{X\sim\mu}\left[\mathcal{S}^\theta g(X)\right]\right)^2$$

is convex in the parameters θ. This follows from standard results on convexity preserving functions as follows. Firstly, due to the linearity of the expectation,

$$\mathbb{E}_{X\sim\mu}\left[\mathcal{S}^\theta g(X)\right] = \langle\theta, \mathbb{E}_{X\sim\mu}\left[\mathcal{T}g(X)\right]\rangle$$

is linear in θ. Then it holds generally that a point wise supremum of linear (convex) functions is convex. In other words, a function $f(x) = \sup_{y\in\mathcal{A}} h(x,y)$, which is convex in x for each y, is convex in x. Lastly, squaring a non-negative convex function preserves convexity. Non-negativity follows by the linearity of the functional $\mathcal{T}$ (and the linearity of $\mathcal{S}^\theta$ respectively), as

$$\mathbb{E}_{X\sim\mu}\left[\mathcal{S}^\theta(-g)(X)\right] = -\mathbb{E}_{X\sim\mu}\left[\mathcal{S}^\theta g(X)\right].$$

Since $g \in \mathcal{H}^d_{\leq 1}$ if and only if $-g \in \mathcal{H}^d_{\leq 1}$, the supremum over $\mathcal{H}^d_{\leq 1}$ is non-negative. □

B.7 Calculations in Example 9

For $b(x) = B^\theta j(x)$, we calculate

$$\begin{aligned}
\langle b(x), g(x)\rangle &= g(x)^T B^\theta j(x) \\
&= \operatorname{tr}\left(g(x)^T B^\theta j(x)\right) \\
&= \operatorname{tr}\left(B^\theta j(x) g(x)^T\right) \\
&= \operatorname{vec}\left((B^\theta)^T\right)^T \operatorname{vec}\left(j(x)g(x)^T\right) \\
&= \langle \operatorname{vec}\left((B^\theta)^T\right), \operatorname{vec}(j(x)\otimes g(x))\rangle,
\end{aligned}$$

where $j(x)\otimes g(x) = j(x)g(x)^T$ denotes the outer product. We defined V to stack the vectorized basis diffusion matrices, i.e., $V(x) = (\operatorname{vec}(v_1(x)v_1(x)^T) \mid \ldots \mid \operatorname{vec}(v_m(x)v_m(x)^T))^T$. Then

$$\begin{aligned}
\operatorname{tr}(a(x)\nabla_x g(x)) &= \operatorname{tr}\left(\sum_{i=1}^m A_i^\theta\, v_i(x)v_i(x)^T\nabla_x g(x)\right) \\
&= \sum_{i=1}^m A_i^\theta \operatorname{tr}\left((v_i(x)v_i(x)^T\nabla_x g(x)\right) \\
&= \sum_{i=1}^m A_i^\theta \operatorname{vec}\left(v_i(x)v_i(x)^T\right)^T \operatorname{vec}\left(\nabla_x g(x)\right) \\
&= \langle \left(A_i^\theta\right)_{i\in[m]}, V(x)\operatorname{vec}\left(\nabla_x g(x)\right)\rangle.
\end{aligned}$$

B.8 Proof of Proposition 10

From Example 9 we have

$$\mathcal{T}g(x) = \begin{pmatrix}\operatorname{vec}\left(j(x)\otimes g(x)\right) \\ V(x)\operatorname{vec}\left(\nabla g(x)\right)\end{pmatrix} \in \Gamma(\mathbb{R}^d, \mathbb{R}^{ld+m}), \tag{65}$$

where V stacks the vectorized basis diffusion matrices, i.e., $V(x) = (\operatorname{vec}(v_1(x)v_1(x)^T) \mid \ldots \mid \operatorname{vec}(v_m(x)v_m(x)^T))^T$. We let $\mathcal{T}^{(i)}$ denote the projection onto the i-th entry and $\mathcal{T}_1$ indicate the application to the first argument. The operator $\mathcal{T}$, in an abuse of notation, is expanded to act on $C^1(\mathbb{R}^d, \mathbb{R}^{d\times d})$ as in Definition 1 defined for the SKDS operator $\mathcal{S}$. We define the following square matrices for the scope of this section;

$$R := \left(\langle \mathbb{E}_{X\sim\mu}\left[(\mathcal{T}_1^{(i)}K(X,\cdot))^T\right], \mathbb{E}_{X'\sim\mu}\left[(\mathcal{T}_1^{(j)}K(X',\cdot))^T\right]\rangle_{\mathcal{H}^d}\right)_{i,j\in[dl+m]} \tag{66}$$

and define

$$M(x,y) := \mathcal{T}_2\left((\mathcal{T}_1 K(x,y))^T\right) \tag{67}$$

as the consecutive application of $\mathcal{T}$ to both arguments of a matrix-valued kernel K. For notational convenience, we sometimes omit the arguments of $M(x,y)$ and only write M. The proof of Proposition 10 further requires the following two lemmas. The next result shows that in R the expectation and the inner product of $\mathcal{H}^d$ commute.

Lemma 22. *Let $\mathcal{T}$ be defined as in Example 9, and let the basis functions j, $v_i v_i^T$ be bounded and continuous. In specific, $\|j\|_\infty$, $\left\|v_i v_i^T\right\|_\infty \leq C$ for all $i \in [m]$, and a $C \in [1, \infty)$. Moreover, let $K = kI_d$ with $k \in C_b^{(2,2)}(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$. Then $\mathbb{E}_{X\sim\mu}[\mathcal{T}g]$ is a continuous linear functional on $\mathcal{H}^d$, and it holds that*

$$R = \mathbb{E}_{X\sim\mu}\left[\mathbb{E}_{Y\sim\mu}\left[\mathcal{T}_2\left((\mathcal{T}_1 K(X,Y))^T\right)\right]\right] = \mathbb{E}_{X\sim\mu}\left[\mathbb{E}_{Y\sim\mu}\left[M(X,Y)\right]\right]. \tag{68}$$

Proof. The linearity follows immediately from the linearity of both the expectation and $\mathcal{T}$. We further show that the operator $\mathcal{T}$ is bounded on $\mathcal{H}^d_{\leq 1}$ as follows. The first component $\operatorname{vec}\left(j(x) \otimes g(x)\right)$ is bounded due to the boundedness assumptions on j and $k(x,x)$ using (43) since $\|j\|_\infty \|g\|_\infty \leq C \sup_{x\in\mathbb{R}^d} k(x,x)$. Boundedness of the second component $V(x)\operatorname{vec}\left(\nabla g(x)\right)$ follows similarly from the boundedness of V and $\partial_{x_j} k(x,x)$ using (44) as $\|v\|_\infty^2 \|\nabla g\|_\infty \leq C^2 \sup_{x\in\mathbb{R}^d} \sqrt{\partial_{x_j} k(x,x)}$. As $\|\mathbb{E}_{X\sim\mu}[\mathcal{T}g]\|_2 \leq (dl+m)\|\mathcal{T}g\|_\infty$, we obtain that $\mathbb{E}_{X\sim\mu}[\mathcal{T}g]$ is bounded on $\mathcal{H}^d_{\leq 1}$, and hence continuous.

By Riesz representation theorem, there exists $(g_{\mu,\mathcal{T}^{(i)}})_{i\in[dl+m]} \subset \mathcal{H}^d$ such that $\mathbb{E}_{X\sim\mu}[\mathcal{T}^{(i)}g] = \langle g_{\mu,\mathcal{T}^{(i)}}, g\rangle_{\mathcal{H}^d}$. Following the same steps as in Appendix B.1 we conclude

$$g_{\mu,\mathcal{T}^{(i)}}(\cdot) = \mathbb{E}_{X\sim\mu}[\mathcal{T}_1^{(i)} K(X,\cdot)^T]. \tag{69}$$

Hence, $R_{ij} = \mathbb{E}_{X\sim\mu}\left[\mathcal{T}^{(i)}\left(\mathbb{E}_{X'\sim\mu}\left[(\mathcal{T}_1^{(j)} K(X',\cdot))^T\right]\right)\right]$. Since $\mathcal{T}$ is a continuous linear operator, $\mathbb{E}$ and $\mathcal{T}$ commute, and the statement follows. □

The following lemma bounds the entries of M.

Lemma 23. *Under the assumptions of Lemma 22, for all $x, y \in \mathbb{R}^d$ it holds that*

$$\|M(x,y)\|_F \leq L_1 := (dl+m)^2 d^2 C^2 C_k, \tag{70}$$

$$m_2(M(x,y)) := \max\left(\left\|\mathbb{E}[MM^T]\right\|_F, \left\|\mathbb{E}[M^T M]\right\|_F\right) \leq L_2 := (dl+m)^3 (d^2 C^2 C_k)^2, \tag{71}$$

where $m_2(M(x,y))$ denotes the per-sample second moment and C_k is a constant dependent on k.

Proof. To bound $\|M(x,y)\|_F$, we explicitly write down $M(x,y) = \mathcal{T}_2\left((\mathcal{T}_1 K(x,y))^T\right)$ as

$$\mathcal{T}_1 K(x,y) = (\mathcal{T}_1 k(x,y)e_1 \mid \ldots \mid \mathcal{T}_1 k(x,y) e_d) \quad \text{with} \quad \mathcal{T}_1 k(x,y) e_i = \begin{pmatrix} \operatorname{vec}\left(j(x) \otimes k(x,y) e_i\right) \\ V(x)\operatorname{vec}\left(e_i \otimes \nabla_x k(x,y)\right) \end{pmatrix}.$$

The transpose of the full matrix is then

$$\begin{aligned}(\mathcal{T}_1 K(x,y))^T &= \left[\begin{array}{cccc|ccc} k(x,y)j(x)^T & 0 & \cdots & 0 & (v_1 v_1^T)_{1\cdot}\nabla_x k & \cdots & (v_m v_m^T)_{1\cdot}\nabla_x k \\ 0 & k(x,y)j(x)^T & \cdots & 0 & (v_1 v_1^T)_{2\cdot}\nabla_x k & \cdots & (v_m v_m^T)_{2\cdot}\nabla_x k \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & k(x,y)j(x)^T & (v_1 v_1^T)_{d\cdot}\nabla_x k & \cdots & (v_m v_m^T)_{d\cdot}\nabla_x k \end{array}\right] \\ &= \left[\begin{array}{cccc|c} k(x,y)j(x)^T & 0 & \cdots & 0 & \\ 0 & k(x,y)j(x)^T & \cdots & 0 & \\ \vdots & \vdots & \ddots & \vdots & v_1 v_1^T \nabla_x k \mid \ldots \mid v_m v_m^T \nabla_x k \\ 0 & 0 & \cdots & k(x,y)j(x)^T & \end{array}\right].\end{aligned}$$

When applying $\mathcal{T}_2$, there are two cases to distinguish. First, calculate

$$\begin{aligned}\mathcal{T}_2(k(x,y)j(x)_i e_j) &= \begin{pmatrix} \operatorname{vec}\left(j(y) \otimes (k(x,y)j(x)_i e_j)\right) \\ V(y)\cdot\operatorname{vec}\left(\nabla_y (k(x,y)j(x)_i e_j)\right) \end{pmatrix} \\ &= j(x)_i \begin{pmatrix} k(x,y)\operatorname{vec}\left(j(y)\otimes e_j\right) \\ V(y)\cdot\operatorname{vec}\left(e_j \otimes \nabla_y k(x,y)\right) \end{pmatrix},\end{aligned}$$

and for the second case

$$\begin{aligned}\mathcal{T}_2(v_i(x)v_i(x)^T\nabla_x k(x,y)) &= \begin{pmatrix}\operatorname{vec}\left(j(y)\otimes(v_i(x)v_i(x)^T\nabla_x k(x,y))\right)\\ V(y)\cdot\operatorname{vec}\left(\nabla_y(v_i(x)v_i(x)^T\nabla_x k(x,y))\right)\end{pmatrix}\\ &= \begin{pmatrix}\operatorname{vec}\left(j(y)\otimes(v_i(x)v_i(x)^T\nabla_x k(x,y))\right)\\ V(y)\cdot\operatorname{vec}\left(v_i(x)v_i(x)^T\nabla_y\nabla_x k(x,y)\right)\end{pmatrix}.\end{aligned}$$

We bound every entry to obtain

$$\left\|j(x)_i k(x,y)\operatorname{vec}\left(j(y)\otimes e_j\right)\right\|_\infty \le C^2 \sup_{x,y\in\mathbb{R}^d} k(x,y), \tag{72}$$

$$\left\|j(x)_i V(y)\cdot\operatorname{vec}\left(e_j\otimes\nabla_y k(x,y)\right\|_\infty \le dC^2 \sup_{x,y\in\mathbb{R}^d}\left\|\nabla_y k(x,y)\right\|_\infty, \tag{73}$$

$$\left\|\operatorname{vec}\left(j(y)\otimes(v_i(x)v_i(x)^T\nabla_x k(x,y))\right)\right\|_\infty \le dC^2 \sup_{x,y\in\mathbb{R}^d}\left\|\nabla_x k(x,y)\right\|_\infty, \tag{74}$$

$$\left\|V(y)\cdot\operatorname{vec}\left(v_i(x)v_i(x)^T\nabla_y\nabla_x k(x,y)\right)\right\|_\infty \le d^2C^2 \sup_{x,y\in\mathbb{R}^d}\left\|\nabla_y\nabla_x k(x,y)\right\|_\infty. \tag{75}$$

We define $C_k := \sup_{x,y\in\mathbb{R}^d}\max(k(x,y), \|\nabla_x k(x,y)\|_\infty, \|\nabla_y k(x,y)\|_\infty, \|\nabla_y\nabla_x k(x,y)\|_\infty)$ and bound the Frobenius norm of $M(x,y)$ with the entrywise bound for each of the $(dl+m)^2$ entries. Then, it holds that

$$\|M(x,y)\|_F \le (dl+m)^2 d^2C^2C_k = L_1 \tag{76}$$

for all x, $y\in\mathbb{R}^d$. We further bound the per-sample second moment by

$$m_2(M(x,y)) \le (dl+m)^3(d^2C^2C_k)^2 = L_2. \tag{77}$$

□

Proof of Proposition 10. By Lemma 5 and substituting the linear parametrization, we have

$$\begin{aligned}\operatorname{SKDS}(\mathcal{S}^\theta,\mu;\mathcal{H}^d_{\le 1}) &= \mathbb{E}_{X\sim\mu}\left[\mathbb{E}_{Y\sim\mu}\left[\mathcal{S}_1^\theta\mathcal{S}_2^\theta K(X,Y)\right]\right]\\ &= \mathbb{E}_{X\sim\mu}\left[\mathbb{E}_{Y\sim\mu}\left[\langle\theta,\mathcal{T}_1\left(\langle\theta,\mathcal{T}_2(K(X,Y))^T\right)\rangle\right]\right]\\ &= \langle\theta,\mathbb{E}_{X\sim\mu}\left[\mathbb{E}_{Y\sim\mu}\left[\mathcal{T}_1\left((\mathcal{T}_2K(x,y))^T\right)\right]\right]\theta\rangle\\ &= \langle\theta,R\theta\rangle.\end{aligned} \tag{78}$$

Note that R is a Gram-matrix, hence PSD. This is an alternative proof for the convexity of the SKDS, although with stronger regularity assumptions.

Similarly, the empirical estimator in (18) can be written using the identity

$$\theta^T M(x,y)\theta = \frac{1}{2}\theta^T\left(M(x,y)+M(x,y)^T\right)\theta,$$

which holds for any real vector θ and square matrix, yielding

$$\begin{aligned}\widehat{\operatorname{SKDS}}(\mathcal{S}^\theta,D;\mathcal{H}^d_{\le 1}) &= \frac{1}{\lfloor N/2\rfloor}\sum_{n=1}^{\lfloor N/2\rfloor}\mathcal{S}_1^\theta\mathcal{S}_2^\theta K(x_{2n-1},x_{2n})\\ &= \langle\theta,\frac{1}{\lfloor N/2\rfloor}\sum_{n=1}^{\lfloor N/2\rfloor}\mathcal{T}_1\left((\mathcal{T}_2K(x_{2n-1},x_{2n}))^T\right)\theta\rangle\\ &= \langle\theta,\frac{1}{\lfloor N/2\rfloor}\sum_{n=1}^{\lfloor N/2\rfloor}\frac{1}{2}\left(\mathcal{T}_1\left((\mathcal{T}_2K(x_{2n-1},x_{2n}))^T\right)+\mathcal{T}_1\left((\mathcal{T}_2K(x_{2n-1},x_{2n}))^T\right)^T\right)\theta\rangle\\ &=: \langle\theta,\frac{1}{\lfloor N/2\rfloor}\sum_{n=1}^{\lfloor N/2\rfloor}R_n\theta\rangle.\end{aligned} \tag{79}$$

Note that R_n are independent copies of $\frac{1}{2}(M + M^T)$. From Lemma 22, we know $\mathbb{E}[M] = R$, and as $R = R^T$, we have $\mathbb{E}[M^T] = \mathbb{E}[M]^T = R$ as well. We conclude

$$\mathbb{E}[R_n] = \mathbb{E}[\frac{1}{2}(M(x_{2n-1}, x_{2n}) + M(x_{2n-1}, x_{2n})^T)] = R,$$

with $(R_n)_{n\in\lfloor N/2\rfloor}$ independent. Using Lemma 23, we apply the error estimation for matrix samples from Tropp (2015, Cor 6.2.1) and get for any $\epsilon > 0$ that

$$\mathbb{P}\left[\left\|\frac{1}{\lfloor N/2\rfloor}\sum_{n=1}^{\lfloor N/2\rfloor} R_n - R\right\|_F \geq \epsilon\right] \leq 2(dl+m)\exp\left(-\frac{\lfloor N/2\rfloor\epsilon^2/2}{L_2 + 2L_1\epsilon/3}\right), \tag{80}$$

for constants L_1 and L_2 as specified in Lemma 23. The Frobenius norm can be lower bounded with basic linear algebra and Weyl's inequality for spectral stability, yielding

$$\begin{aligned}\left\|\frac{1}{\lfloor N/2\rfloor}\sum_{n=1}^{\lfloor N/2\rfloor} R_n - R\right\|_F &\geq \left\|\frac{1}{\lfloor N/2\rfloor}\sum_{n=1}^{\lfloor N/2\rfloor} R_n - R\right\|_{\text{op},2} \\ &\geq \max_{k\in[dl+m]}\left|\lambda_{(k)}\left(\frac{1}{\lfloor N/2\rfloor}\sum_{n=1}^{\lfloor N/2\rfloor} R_n\right) - \lambda_{(k)}(R)\right|,\end{aligned} \tag{81}$$

where $\lambda_{(k)}(\cdot)$ denotes the k-th eigenvalue. Note that Weyl's inequality is only applicable to symmetric (or hermitian) matrices. We argued that R is PSD and, in specific, symmetric. The estimation matrix R_n is constructed to be symmetric as well. We conclude from (81) and the positive semi-definiteness of R that the empirical estimator $\hat{\text{SKDS}}(\mathcal{S}^\theta, D; \mathcal{H}^d_{\leq 1})$ is ϵ-strongly quasiconvex with high probability, In other words, it holds that

$$\hat{\text{SKDS}}(\mathcal{S}^\theta, D; \mathcal{H}^d_{\leq 1}) + \epsilon\left\|\theta\right\|_2 \tag{82}$$

is convex in θ with probability greater than $1 - 2(dl+m)\exp\left(-\frac{\lfloor N/2\rfloor\epsilon^2/2}{L_2+2L_1\epsilon/3}\right)$. The result follows from rearranging for N. □

C EXPERIMENTS

C.1 Kernel Choices

In this work, we use the following kernels for model construction and benchmarking. Each kernel emphasizes different properties of the input space or adjusts standard kernels to better suit our setting.

1. **RBF Kernel:**

$$k_{\text{RBF}}(x, y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right).$$

   This standard Gaussian (Radial Basis Function) kernel promotes smoothness by measuring similarity based on the Euclidean distance between inputs. It is widely used for its universal approximation properties and strong empirical performance.

2. **Tilted RBF Kernel:**

$$k_{\text{RBF, tilted}}(x, y) = \frac{1}{w_1(x)w_1(y)}\exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right).$$

   This variant of the RBF kernel downweights regions far from the origin using the weight function $w_1(x) = (1 + \|x\|_2^2)^{1/2}$. It effectively biases the kernel toward the origin and suppresses the influence of distant samples, which can be beneficial in high-dimensional or heavy-tailed settings.

3. **IMQ+ Kernel:**

$$k_{\mathrm{IMQ}^+}(x,y) = \frac{1}{w_1(x)w_1(y)}\left(\frac{1}{w_1(x-y)} + (1+\langle x,y\rangle)\right),$$

where $w_1(x) = (1+\|x\|_2^2)^{1/2}$. This kernel combines a repulsive inverse multiquadric term and an attractive linear term, both scaled by polynomially decaying weights. It was proposed in Kanagawa et al. (2022, Corollary 3.4) and is designed to control the spectral decay of the associated integral operator, improving generalization in certain distributional regimes.

C.2 Compute Infrastructure

All experiments were conducted on a high-performance computing (HPC) cluster equipped with *Intel(R) Xeon(R) Platinum 8480+ (Sapphire Rapids)* processors.

Jobs were scheduled using `SLURM` in combination with the `jobfarm` framework to parallelize and manage large batches of experiments. A typical job allocation requested `2` nodes with `200` tasks (`2` CPUs per task) and a maximum runtime of 24 hours. The runtime environment was managed with `Conda`, and all dependencies were contained in a dedicated environment to ensure reproducibility.

For transparency, the following example job script illustrates the scheduling setup:

```
#!/bin/bash
#SBATCH -J JobFarm
#SBATCH --nodes=2
#SBATCH --ntasks=200
#SBATCH --cpus-per-task=2
#SBATCH --time=24:00:00
...
jobfarm start experiment/command_list.txt
```

This configuration enabled efficient distribution of workloads across multiple CPUs while maintaining consistent software environments.

C.3 Baselines

We compare against several established methods for causal discovery:

- **GIES** (Hauser and Bühlmann, 2012), implemented via the Causal Discovery Toolbox (MIT License)[4].
- **IGSP** (Wang et al., 2017), using the CausalDAG package (3-Clause BSD License)[5].
- **DCDI** (Brouillard et al., 2020), with the authors' Python implementation (MIT License).
- **NODAGS** (Sethuraman et al., 2023), using the official implementation (Apache 2.0 License).
- **LLC** (Hyttinen et al., 2012), based on the NODAGS code (Apache 2.0 License) with extensions by Lorch et al. (2024) with code published under MIT License (Lorch, 2024).
- **KDS** (Lorch, 2024), using the official repository (MIT License)[6].

C.4 Results

To evaluate whether our method is significantly better than a baseline method by a relative margin of 5%, we perform a paired Wilcoxon signed-rank test on the log-transformed metric ratios. In specific, for any method and data generation process we calculate the Wasserstein-2 distance and mean squared error (MSE) in every dimension. For one of our methods and a baseline method we define a paired array $(\mathrm{our}_i, \mathrm{baseline}_i)_{i\in[50*10]}$ of

[4] https://github.com/FenTechSolutions/CausalDiscoveryToolbox
[5] https://github.com/uhlerlab/causaldag
[6] https://github.com/larslorch/stadion

metric values, where 50 is the amount of generated datasets and 10 is the number of test interventions per dataset. By using the log transform, multiplicative differences become additive. Let $r_i = \log(\text{our}_i) - \log(\text{baseline}_i)$. A 5% improvement (ours is 5% smaller) means $\text{our}_i \leq 0.95 \cdot \text{baseline}_i$, which is equivalent to $\log(\text{our}_i) - \log(\text{baseline}_i) \leq \log(0.95)$. So we test

$$H_0 : \text{median}(r_i) = 0 \quad \text{vs} \quad H_1 : \text{median}(r_i) < \log(0.95).$$

Equivalently, we can define

$$d_i = r_i - \log(0.95)$$

and run a Wilcoxon signed-rank test on d_i versus 0 (i.e., test whether $\text{median}(d_i) < 0$). We run this as well with roles of our method and the baseline method switched. The results for both of our methods are reported in Table 3 and Table 4.

As detailed in Table 3, **SKDS (Linear)** achieves statistically significant improvements (a relative margin of at least 5%) over most baselines across all datasets, while performing on par with **KDS**. **SKDS (MLP)** also performs strongly and is competitive with all baselines, but shows significant losses in purely linear settings, where simpler models are inherently better suited. This reflects a natural trade-off between flexibility and inference difficulty: more expressive models adapt well to nonlinear data but may underperform when the ground truth is linear.

In the significance tests, the **SKDS** variants are only outperformed by the baseline **KDS** and only for one data-generation setting (linear SDEs with Erdős–Rényi structure) when evaluated with the Wasserstein-2 distance. This finding aligns with Proposition 6, which establishes that the SKDS objective can be upper bounded by the Wasserstein-2 distance. In this sense, the observed gap may reflect the tighter sensitivity of Wasserstein-2 distance to distributional differences beyond the mean, whereas MSE primarily emphasizes mean accuracy.

Baseline / Dataset	SDE-ER	SDE-SF	SCM-ER	SCM-SF	SERGIO-ER	SERGIO-SF
GIES	/ *	/ *	/ *	/	* / *	* / *
IGSP	* / *	/ *	* / *	* / *	* / *	* / *
DCDI	* / *	* / *	* / *	* / *	* /	/
LLC	* / *	/ *	* / *	* / *	* / *	* / *
NODAGS	* / *	/ *	* / *	/ *	* / *	/
KDS (Linear)	/	/	* / *	* / *	* / *	* / *
KDS (MLP)	* /	* / *	* / *	/	/	/

Table 3: Significance tests for **SKDS (Linear)**. Black star indicates our method significantly outperforms the baseline; red star indicates baseline significantly outperforms ours. Each cell shows MSE / Wasserstein results.

Baseline / Dataset	SDE-ER	SDE-SF	SCM-ER	SCM-SF	SERGIO-ER	SERGIO-SF
GIES	* /	/ *	* /	* / *	* / *	* / *
IGSP	* / *	/ *	* / *	/ *	* / *	* / *
DCDI	/ *	* / *	* / *	* / *	/ *	/ *
LLC	* / *	/ *	* / *	/ *	* / *	* / *
NODAGS	* / *	/ *	* / *	/ *	/ *	/ *
KDS (Linear)	* / *	* /	* / *	* / *	* / *	* / *
KDS (MLP)	* / *	/	/	/	* /	/

Table 4: Significance tests for **SKDS (MLP)**. Black star indicates our method significantly significantly outperforms the baseline by at least 5%; red star indicates baseline significantly outperforms ours. Each cell shows MSE / Wasserstein results.

The results in Fig. 4 provide a complementary perspective on model performance. They show that the choice of causal structure does not substantially affect the comparative behavior of the methods. In the linear setting, all approaches achieve similar accuracy with respect to MSE. By contrast, in the gene expression experiments, only **SKDS (Linear)** remains competitive as a linear model, whereas approaches with nonlinear drift functions exhibit a clear advantage.

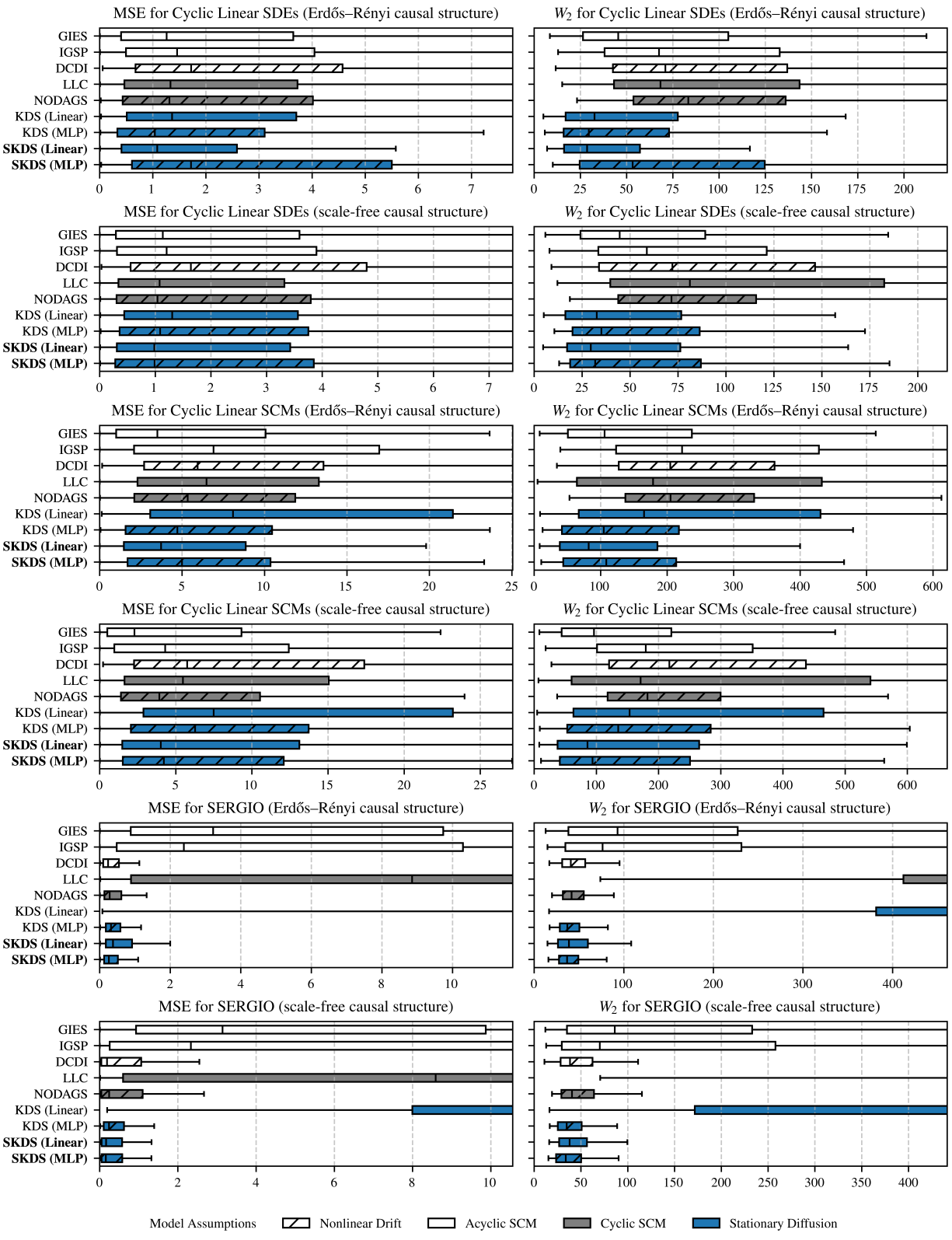

Figure 4: Benchmarking results for $d = 20$ variables. The MSE (left column) and the Wasserstein-2 distance (right column) distance MSE were computed from 10 test interventions on unseen target variables across 50 randomly generated systems and 6 different data generating methods (rows). Box plots depict the medians and interquartile ranges (IQR), with whiskers extending to the largest value within 1.5 times the IQR from the boxes.

D Code

The Stein-type kernel deviation from stationarity can be used as a drop-in replacement for the framework by Lorch (2024). The critical code snippet is to be found in Fig. 5.

```
def stein_type_kds_operator(h, argnum, hdim):
    def h_out(x, y, *args):
        assert x.ndim == y.ndim == 1
        assert x.shape == y.shape
        z = x if argnum == 0 else y
        f_x = f(z, *args)
        sigma_x = sigma(z, *args)
        grad_h = jax.jacfwd(h, argnums=argnum)(x, y, *args)
        if hdim == 0:
            return h(x,y,*args) * f_x + 0.5 * sigma_x.T @ sigma_x @ grad_h
        if hdim == 1:
            return f_x @ h(x,y,*args) + 0.5 * jnp.trace(sigma_x @ sigma_x.T @ grad_h)
    return h_out

def loss_term(x, y, *args):
    return stein_type_kds_operator(
        stein_type_kds_operator(_kernel, 0, 0), 1, 1
    )(x, y, *args)
```

Figure 5: Nested Stein-type KDS operator defining the custom objective function integrated into the `stadion` framework (Lorch, 2024).