Stein Transport for Generative Data Augmentation

Clémentine Chazal*
CREST, ENSAE, IP Paris
Palaiseau, France
clementine.chazal@ensae.fr

Linfeng Wang*
Department of Mathematics
King's College London
London, United Kingdom
linfeng.2.wang@kcl.ac.uk

Nikolas Nüsken
Department of Mathematics
King's College London
London, United Kingdom
nikolas.nusken@kcl.ac.uk

Anna Korba
CREST, ENSAE, IP Paris
Palaiseau, France
anna.korba@ensae.fr

Abstract

Generating additional samples from limited observations is important when data acquisition or Monte Carlo simulation is costly. We propose a nonparametric approach to generative modeling based on Stein operators and probability transport that learns from target samples alone. Using an Ornstein–Uhlenbeck probability path, we derive a tractable conditional regression objective for learning the transport field, without evaluating the target density or its gradient. Restricting the field to a reproducing kernel Hilbert space yields closed-form kernel ridge regression solutions at each time step, with deterministic, stochastic, and coordinatewise variants. We establish convergence rates for the excess conditional Stein risk and show that the regression targets are noise-free given the augmented inputs, identifying a structural distinction from denoising score matching. Experiments on Gaussian mixtures and Bayesian posterior emulation show effective generation in small-data regimes and greater robustness to hyperparameter choices than kernel score matching. On tabular benchmarks, we obtain competitive downstream predictive performance and explore the use of generated samples for data augmentation.

1 Introduction

Generative modeling aims to produce new samples from a target probability distribution given a finite set of observations. Recent progress in this area has been dominated by neural approaches such as diffusion models (Ho et al., 2020; Song et al., 2021) or flow matching (Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2025). However, generative modeling is often designed and studied in large-data regimes, where highly expressive neural architectures can be trained effectively. While very successful in large-scale settings such as image generation, these methods often require substantial training effort, careful architectural choices, and large datasets, which can limit their applicability in alternative scientific and statistical settings (Kotelnikov et al., 2023).

In this work, we focus on a different regime: generative modeling where the available number of samples from the target distribution is small or moderate. This situation arises naturally in scientific applications where data acquisition is expensive. For instance, one may wish to augment a limited dataset of observations in order to improve the stability or generalisation of predictive models (Wang et al., 2025). This has recently led to the development of generative data augmentation,

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

which scales datasets by obtaining synthetic labeled examples from a trained conditional generative model (Yamaguchi et al., 2020; Zheng et al., 2023; Trabucco et al., 2024). In Bayesian inference, generating posterior samples with Markov Chain Monte Carlo (MCMC) (Roberts and Rosenthal, 2004) can be costly, motivating the use of generative models to produce additional samples.

We propose a kernel-based approach to generative modeling using Stein transport and probability paths. Our aim is to learn a reusable sampler from a limited dataset, enabling the generation of additional synthetic observations for downstream tasks. The kernel formulation provides explicit regularisation and reduces estimation of the transport field to solving linear systems at each time step. Our method is inspired by recent developments in Stein transport for Bayesian inference (Nüsken, 2024), where interacting particle systems are used to transform a reference distribution into a target distribution by solving a continuity equation driven by a Stein operator. Previously developed Stein transport methods (Nüsken, 2024; Liu and Wang, 2016) typically assume access to the target density or its gradient, which is standard in a Bayesian inference setting, but which makes them unsuitable for purely sample-based settings.

Conceptually, our method learns a transport field along an Ornstein–Uhlenbeck probability path, whose forward dynamics progressively transform the target distribution towards a Gaussian reference. We fit the action of the conditional Stein operator on this field to an explicit target derived from the path's Gaussian conditional densities. This yields a tractable regression objective using only target samples and added Gaussian noise. Whereas score matching estimates the marginal space score (Song et al., 2021), our formulation learns the transport field through a regression involving its divergence and the conditional space score.

Contributions. Our main contributions are as follows:

- We introduce a Stein transport framework for generative modeling from target samples alone. Using an Ornstein–Uhlenbeck probability path, we derive a tractable conditional regression objective for learning the transport field.
- We derive closed-form estimators of the transport field in a reproducing kernel Hilbert space and develop deterministic and stochastic sampling algorithms. We also introduce vectorised variants that fit each component of the field through a separate scalar regression.
- We establish convergence rates for the excess conditional Stein risk at each fixed time. We also show that the regression targets are noise-free given the augmented inputs comprising the original and noised samples, and discuss this property in relation to denoising score matching.
- We evaluate the methods on Gaussian mixtures, Bayesian posterior emulation, and tabular data augmentation. The Gaussian mixture and posterior experiments show greater robustness to hyperparameter choices than kernel score matching. On tabular benchmarks, we obtain competitive downstream predictive performance and investigate the benefits of augmenting real data with generated samples.

The remainder of the paper presents the diffusion background and Stein transport formulation (Sections 2 and 3), followed by the kernel algorithms (Section 4), theoretical analysis (Section 5), and numerical experiments (Section 6).

Notation. Let $\mathcal{P}(\mathbb{R}^d)$ denote the set of probability measures on $\mathbb{R}^d$. We denote by γ_d the standard Gaussian on $\mathbb{R}^d$. For any $\pi \in \mathcal{P}(\mathbb{R}^d)$, $L^2(\pi)$ is the Hilbert space of (equivalence classes of) functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\int |f|^2 \, \mathrm{d}\pi < \infty$. We denote by $\| \cdot \|_{L^2(\pi)}$ and $\langle \cdot, \cdot \rangle_{L^2(\pi)}$ the norm and the inner product of $L^2(\pi)$. For vector-valued functions $g : \mathbb{R}^d \to \mathbb{R}^p$, we abuse the notation of $L^2(\pi)$ and claim $g \in L^2(\pi)$ if $g_i \in L^2(\pi)$ for all $i = 1, \ldots, p$ along with $\|g\|^2_{L^2(\pi)} := \sum_{i=1}^p \|g_i\|^2_{L^2(\pi)}$. If $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, we denote by ∇f the gradient of f. For a vector-valued function $g : \mathbb{R}^d \to \mathbb{R}^d$, if g_i is differentiable for all $i = 1, \cdots, d$, $\nabla \cdot g$ denotes the divergence of g. We also denote by Δg the Laplacian of g, where $\Delta g = \nabla \cdot \nabla g$.

2 Background: Ornstein–Uhlenbeck paths and reverse dynamics

Given samples from a probability distribution of interest, the goal of generative modeling is to generate additional samples from the same distribution. A popular approach (Song et al., 2021) is to learn to transform samples from a standard Gaussian distribution into samples from the data distribution, using the time-reversal of an Ornstein-Uhlenbeck process.

Ornstein–Uhlenbeck path. Let $\pi_\star$ denote the target distribution from which the observed data are assumed to be drawn. For a fixed time horizon $T > 0$, we construct a forward probability path $(\vec{\pi}_t)_{t\in[0,T]}$ by evolving samples from $\pi_\star$ according to the Ornstein–Uhlenbeck process

$$\mathrm{d}\vec{X}_t = -\vec{X}_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad \vec{X}_0 \sim \pi_\star, \tag{1}$$

where B_t is a standard Brownian motion in $\mathbb{R}^d$, and we set $\vec{\pi}_t := \mathrm{Law}(\vec{X}_t)$. By construction, this gives $\vec{\pi}_0 = \pi_\star$. Moreover, for each fixed $t \in [0, T]$, conditional on the initial data point $X_\star = \vec{X}_0$, the distribution at time t satisfies

$$\vec{X}_t \mid \vec{X}_0 = X_\star \overset{d}{=} e^{-t}X_\star + \sigma_t\varepsilon, \qquad \varepsilon \sim \gamma_d, \quad \sigma_t^2 = 1 - e^{-2t}. \tag{2}$$

Here, ε is independent of $X_\star$.

Thus the path gradually reduces the contribution of the initial sample while adding Gaussian noise. For sufficiently large T, $\vec{\pi}_T$ is close to γ_d, since the Ornstein–Uhlenbeck process converges exponentially fast to its invariant distribution γ_d (Pavliotis, 2014, Section 4.4).

Reverse dynamics. The time reversal of the Ornstein–Uhlenbeck process in (1) over the interval $[0, T]$ satisfies, for $0 \le t < T$,

$$\mathrm{d}\overleftarrow{X}_t = \Big(\overleftarrow{X}_t + 2\nabla \log \overleftarrow{\pi}_t(\overleftarrow{X}_t)\Big)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \qquad \overleftarrow{X}_0 \sim \vec{\pi}_T, \tag{3}$$

where $\overleftarrow{\pi}_t := \mathrm{Law}(\overleftarrow{X}_t) = \vec{\pi}_{T-t}$ (Haussmann and Pardoux, 1986). As the time reversal of the continuous forward OU process, the exact reverse process has continuous trajectories on $[0, T]$ and terminal law $\pi_\star$. In practice, one instead initialises from Gaussian noise, $\overleftarrow{X}_0 \sim \gamma_d$, using the approximation $\gamma_d \approx \vec{\pi}_T$ for large T. The terminal variable $\overleftarrow{X}_T$ is then approximately distributed according to $\pi_\star$, so the reverse dynamics define a generative model from Gaussian noise to data.

Implementing this model requires discretising (3) in time, for example by Euler–Maruyama, and estimating the scores $\nabla \log \overleftarrow{\pi}_t$. These scores can be learned by score matching (Hyvärinen and Dayan, 2005; Vincent, 2011), typically using neural networks (Song et al., 2021). As an alternative to (3), one may use the deterministic probability flow (Chen et al., 2023), which, when initialised from $\vec{\pi}_T$, has the same time marginals $(\overleftarrow{\pi}_t)_{t\in[0,T]}$ as the reverse SDE and evolves, for $0 \le t < T$, according to the ODE

$$\mathrm{d}\overleftarrow{X}_t = \overleftarrow{v}_t(\overleftarrow{X}_t)\,\mathrm{d}t, \qquad \overleftarrow{v}_t(x) = x + \nabla \log \overleftarrow{\pi}_t(x). \tag{4}$$

After time discretisation, (4) yields another generative process: starting from Gaussian noise at time 0, it produces an approximate sample from $\pi_\star$ at time T.

These constructions reduce generative modeling to learning a vector field whose dynamics reproduce, in reverse order, the marginals of the forward Ornstein–Uhlenbeck process. In the next section, we formulate this learning problem using Stein equations.

3 Learning to transport with Stein equations

The reverse dynamics in Section 2 realise a prescribed probability path. We now formulate the problem of learning such vector fields using Stein equations (Nüsken, 2024).

Stein transport. To present the method in a general context, we consider dynamics of the form

$$\mathrm{d}X_t = v_t(X_t)\,\mathrm{d}t + \sqrt{2\alpha}\,\mathrm{d}W_t, \qquad X_0 \sim \pi_0, \tag{5}$$

Here, $\alpha \geq 0$ controls diffusion: $\alpha = 0$ gives deterministic dynamics, while $\alpha = 1$ recovers the diffusion coefficient in (3). Writing $\rho_t := \mathrm{Law}(X_t)$, we seek v_t such that $\rho_t = \pi_t$ for a prescribed path $(\pi_t)_{t\in[0,T]}$. Its densities must therefore satisfy the Fokker–Planck equation

$$\partial_t \pi_t + \nabla \cdot (\pi_t v_t) = \alpha \Delta \pi_t. \tag{6}$$

Dividing (6) by π_t yields the *Stein equation*

$$S_{\pi_t} v_t = -\partial_t \log \pi_t + \alpha \frac{\Delta \pi_t}{\pi_t}, \tag{7}$$

where the *Stein operator* (Anastasiou et al., 2023) is defined by

$$S_{\pi_t} v := \frac{1}{\pi_t} \nabla \cdot (\pi_t v) = \nabla \log \pi_t \cdot v + \nabla \cdot v, \tag{8}$$

for strictly positive differentiable densities π_t, mapping vector fields to scalar fields. We call $\partial_t \log \pi_t$ the *time score* and $\nabla \log \pi_t$ the *space score*. We learn a driving field v_t by minimising the least-squares objective

$$L_t(v_t) := \mathbb{E}_{X\sim\pi_t} \left[\left(S_{\pi_t} v_t(X) + \partial_t \log \pi_t(X) - \alpha \frac{\Delta \pi_t}{\pi_t}(X) \right)^2 \right]. \tag{9}$$

The following extends Nüsken (2024, Proposition 1) from deterministic dynamics to $\alpha \geq 0$:

Proposition 3.1 (Path matching)**.** *Fix $\alpha \geq 0$ and a sufficiently regular path $(\pi_t)_{t\in[0,T]}$ of strictly positive densities. Assume* (5) *is well defined and its Fokker–Planck equation has a unique solution with initial law π_0. If $L_t(v_t) = 0$ for all $t \in [0,T]$, then $\mathrm{Law}(X_t) = \pi_t$ throughout $[0,T]$.*

The marginal densities and their derivatives are unknown in generative modeling; Section 4 therefore constructs a tractable conditional surrogate along the OU path.

4 Stein transport in the context of generative modeling

In this section, we adapt the Stein transport objective (9) to generative modeling and propose tractable schemes based on kernel methods and ridge regression. We allow $\alpha \geq 0$, covering deterministic ($\alpha = 0$) and stochastic ($\alpha > 0$) dynamics.

Following Section 2, we would like to minimise the Stein transport objective (9) along the reverse OU path $(\overleftarrow{\pi}_t)_{t\in[0,T]}$, connecting an approximate auxiliary Gaussian at $t = 0$ to the data distribution at $t = T$. Recall that this path is induced by the forward dynamics (1) via $\vec{\pi}_t = \mathrm{Law}(\vec{X}_t)$ and time reversal: $\overleftarrow{\pi}_t := \vec{\pi}_{T-t}$.

Learning the backward drift from the forward process. In the deterministic case when trying to solve the Stein equation from the forward process with $\alpha = 0$, the velocity field of the backward process is simply given by $\overleftarrow{v}_t = -\vec{v}_{T-t}$. However, in the stochastic case, given a drift $(\vec{v}_t)_t$ of the forward process, a backward drift is given by $\overleftarrow{v}_t = -\vec{v}_{T-t} + 2\alpha \nabla \log \vec{\pi}_{T-t}$. To avoid having to use the expression $\nabla \log \vec{\pi}_{T-t}$ we instead consider the loss

$$R_t(v_t) := \mathbb{E}_{\vec{X}_t\sim\vec{\pi}_t} \left[\left(S_{\vec{\pi}_t} v_t(\vec{X}_t) + \partial_t \log \vec{\pi}_t(\vec{X}_t) + \alpha \frac{\Delta \vec{\pi}_t}{\vec{\pi}_t}(\vec{X}_t) \right)^2 \right], \tag{10}$$

where the sign before the stochastic term has been changed. This ensures that a drift for the backward process is given by $\overleftarrow{v}_t = -\vec{v}_{T-t}$ where $\vec{v}_t$ is a solution of Equation (10) (this is not a drift for the forward process). The positive sign of the diffusion term follows from the substitutions $\overleftarrow{v}_s = -v_{T-s}$ and $\overleftarrow{\pi}_s = \vec{\pi}_{T-s}$ in (9): the Stein and time-derivative terms change sign, and the overall minus sign disappears upon squaring.

The conditioning trick. Evaluating R_t requires the space score $\nabla \log \vec{\pi}_t$, time score $\partial_t \log \vec{\pi}_t$, and density Laplacian ratio $\Delta \vec{\pi}_t / \vec{\pi}_t$ of the forward marginal, all of which are unknown. However, conditional on $X_\star = \vec{X}_0$, the forward OU distribution is Gaussian by (2). For $t > 0$, the corresponding

conditional quantities are explicit:

$$\nabla_x \log \vec{\pi}_{t|\star}(x|X_\star) = -\frac{x - e^{-t}X_\star}{\sigma_t^2}, \tag{11}$$

$$\partial_t \log \vec{\pi}_{t|\star}(x|X_\star) = -d\frac{e^{-2t}}{\sigma_t^2} - \frac{e^{-t}}{\sigma_t^2}(x - e^{-t}X_\star)\cdot X_\star + \frac{e^{-2t}}{\sigma_t^4}\left\|x - e^{-t}X_\star\right\|^2, \tag{12}$$

$$\frac{\Delta\vec{\pi}_{t|\star}(x|X_\star)}{\vec{\pi}_{t|\star}(x|X_\star)} = \frac{\|x - e^{-t}X_\star\|^2}{\sigma_t^4} - \frac{d}{\sigma_t^2},$$

Here, ∂_t is taken at fixed x, and $\sigma_t^2 = 1 - e^{-2t}$. Define the shorthand

$$\Psi_t(x_t, x_\star) = -\partial_t \log \vec{\pi}_{t|\star}(x_t|x_\star) - \alpha\frac{\Delta\vec{\pi}_{t|\star}(x_t|x_\star)}{\vec{\pi}_{t|\star}(x_t|x_\star)}.$$

Conditioning the forward marginal quantities on $X_\star$ leads to the following surrogate objective.

Proposition 4.1 (Conditional Stein surrogate)**.** *Fix $t \in (0, T]$ and let $(\vec{X}_s)_{s\geq 0}$ be the OU process in* (1)*, with $\vec{\pi}_t = \mathrm{Law}(\vec{X}_t)$. Then, for any sufficiently regular vector field v,*

$$R_t(v) \leq R_t^c(v), \tag{13}$$

where

$$R_t^c(v) := \mathbb{E}_{X_\star\sim\pi_\star,\ \varepsilon_t\sim\gamma_d}\left[\left((S_{\vec{\pi}_{t|\star}}v)(\vec{X}_t|X_\star) - \Psi_t(\vec{X}_t, X_\star)\right)^2\right], \tag{14}$$

with $\vec{X}_t = e^{-t}X_\star + \sigma_t\varepsilon_t$ and with the conditional Stein operator defined as a natural modification of (8)*:*

$$(S_{\vec{\pi}_{t|\star}}v)(x_t|x_\star) := \nabla \log \vec{\pi}_{t|\star}(x_t|x_\star)\cdot v(x_t) + (\nabla\cdot v)(x_t), \tag{15}$$

The proof is deferred to Section A and follows from Jensen's inequality; similar ideas are common in the literature on denoising diffusion models (Vincent, 2011; Song and Ermon, 2019). The conditional objective upper-bounds the marginal Stein risk but need not share its minimisers. For $\alpha = 0$, it corresponds to the fixed-time form of the squared conditional divergence-matching loss of Huang et al. (2025, Appendix D), expressed here using conditional Stein operators. The additional diffusion correction accommodates $\alpha > 0$. The surrogate objective (14) is tractable from samples $X_\star \sim \pi_*$ and Gaussian noise $\varepsilon \sim \gamma_d$, and minimising it encourages the vector field v to push the dynamics along the prescribed path $(\bar{\pi}_t)_{t\in[0,T]}$, according to (13) and Proposition 3.1.

Solution in an RKHS. We now minimise a regularised empirical version of (14) in a vector-valued reproducing kernel Hilbert space (RKHS). Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive definite kernel with continuous first derivatives and continuous mixed derivatives $\partial_{x_i}\partial_{y_j}k$ for all i, j, and let $\mathcal{H}$ be its reproducing kernel Hilbert space, with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$; see Steinwart and Christmann (2008, Chapter 4) for background. We write $\mathcal{H}^d$ for the product RKHS of vector fields $v = (v_1, \ldots, v_d)$ with $v_i \in \mathcal{H}$, equipped with inner product $\langle v, w\rangle_{\mathcal{H}^d} = \sum_{\ell=1}^d \langle v_\ell, w_\ell\rangle_{\mathcal{H}}$. For any fixed $t \in (0, T]$, we construct the *regularised empirical counterpart* of (14):

$$\widehat{R}_{t,\lambda}^c(v) := \frac{1}{n}\sum_{j=1}^n \left((S_{\vec{\pi}_{t|\star}}v)(\vec{X}_t^{(j)}|X_\star^{(j)}) - \Psi_t(\vec{X}_t^{(j)}, X_\star^{(j)})\right)^2 + \lambda\|v\|_{\mathcal{H}^d}^2, \tag{16}$$

where $\lambda > 0$ is a regularisation parameter. The least-squares objective (16) is based on iid samples $(X_\star^{(j)}, \varepsilon^{(j)})_{j=1}^n$ from $\pi_\star \otimes \gamma_d$ and the OU noising relation $\vec{X}_t^{(j)} = e^{-t}X_\star^{(j)} + \sigma_t\varepsilon_t^{(j)}$ from (2). Our approach crucially rests on the fact that the minimiser of (16) can be obtained in closed form:

Proposition 4.2 (Kernel ridge regression for conditional Stein operators)**.** *The (unique) minimiser of* (16) *over $\mathcal{H}^d$ is given by*

$$\hat{v}_{t,\lambda} = \frac{1}{n}\sum_{j=1}^n \phi_j\left(-k(\cdot, \vec{X}_t^{(j)})\frac{\varepsilon^{(j)}}{\sigma_t} + \nabla_2 k(\cdot, \vec{X}_t^{(j)})\right), \tag{17}$$

where the weights $\phi = (\phi_1, \ldots, \phi_n)^\intercal \in \mathbb{R}^n$ are determined by the linear system

$$\left(\frac{1}{n}\Xi_t + \lambda I_n\right)\phi = \mathbf{\Psi}_t. \tag{18}$$

Here, I_n is the $n \times n$ identity matrix, $(\mathbf{\Psi}_t)_j := \Psi_t(\vec{X}_t^{(j)}, X_\star^{(j)})$ for $j = 1, \dots, n$, and $\Xi_t \in \mathbb{R}^{n \times n}$ has entries

$$(\Xi_t)_{ij} = k(\vec{X}_t^{(i)}, \vec{X}_t^{(j)}) \frac{\varepsilon^{(i)} \cdot \varepsilon^{(j)}}{\sigma_t^2} + \nabla_1 \cdot \nabla_2 k(\vec{X}_t^{(i)}, \vec{X}_t^{(j)}) - \frac{\varepsilon^{(i)}}{\sigma_t} \cdot \nabla_2 k(\vec{X}_t^{(i)}, \vec{X}_t^{(j)}) - \frac{\varepsilon^{(j)}}{\sigma_t} \cdot \nabla_1 k(\vec{X}_t^{(i)}, \vec{X}_t^{(j)}).$$

The proof uses the representer theorem (Wahba, 1990; Schölkopf et al., 2001); see Section B.1.

We are now ready to combine the OU paths from Section 2 with the path matching result from Proposition 3.1 and the exact solution to the regularised objective from Proposition 4.2, finally arriving at Algorithm 1. We use an Euler–Maruyama discretisation of backward dynamics of the form (5), using the velocity field (17), estimated at backward time t_i from noised data samples $\vec{X}_{T-t_i}^{(j)}$ at the corresponding forward noising time $T - t_i$.

Algorithm 1 Stein Transport (ST)

Require: time grid $\{0 = t_0 < t_1 < \cdots < t_m = T\}$, training samples $(X_\star^{(j)})_{j=1}^n$, number of generated samples p, regularisation parameter $\lambda > 0$, stochastic parameter $\alpha \geq 0$.

Ensure: synthetic samples $(\overleftarrow{X}_T^{(r)})_{r=1}^p$.

1: Initialize $\overleftarrow{X}_0^{(r)} \sim \gamma_d$ for $r = 1, \dots, p$.
2: **for** $i = 0$ to $m - 1$ **do**
3: Simulate noised training samples $\vec{X}_{T-t_i}^{(j)} = X_\star^{(j)} e^{-(T-t_i)} + \sigma_{T-t_i} \varepsilon_{T-t_i}^{(j)}$, $\varepsilon_{T-t_i}^{(j)} \sim \gamma_d$, for $j = 1, \dots, n$.
4: Assemble and solve the linear system (18) at forward time $T - t_i$, return $\hat{v}_{T-t_i,\lambda}$ using (17).
5: Evolve synthetic samples: $\overleftarrow{X}_{t_{i+1}}^{(r)} = \overleftarrow{X}_{t_i}^{(r)} - (t_{i+1} - t_i)\hat{v}_{T-t_i,\lambda}(\overleftarrow{X}_{t_i}^{(r)}) + \sqrt{2\alpha(t_{i+1} - t_i)}\eta_{t_i}^{(r)}$, $\eta_{t_i}^{(r)} \sim \gamma_d$.
6: **end for**

Vectorised form. We further adopt a vectorisation strategy similar to that of Yu et al. (2025, Section 3.2), who propose a vectorised formulation of their learning objective. In particular, the original surrogate objective derived in Proposition 4.1 can be rewritten as

$$R_t^c(v) = \mathbb{E}_{X_\star, \varepsilon_t}\left[\left(\sum_{m=1}^d \Big((S_{\vec{\pi}_{t|\star}}^m v^m)(\vec{X}_t, X_\star) - \Psi_t{}^m(\vec{X}_t, X_\star)\Big)\right)^2\right],$$

where v^m denotes the m-th coordinate of v, and we have the coordinate-wise contributions

$$S_{\vec{\pi}_{t|\star}}^m v^m(\vec{X}_t, X_\star) = (\partial_m v^m)(\vec{X}_t) - \frac{\varepsilon_t^m}{\sigma_t} v^m(\vec{X}_t),$$
$$\Psi_t{}^m(\vec{X}_t, X_\star) = \frac{e^{-t}}{\sigma_t} X_\star^m \varepsilon_t^m + \frac{e^{-2t} + \alpha}{\sigma_t^2}\big(1 - (\varepsilon_t^m)^2\big), \qquad m = 1, \dots, d.$$

Here, $X_\star^m$ and ε_t^m denote the m-th components of the vectors $X_\star$ and ε_t, respectively, and $\partial_m v^m := \frac{\partial v^m}{\partial x^m}$ denotes the m-th partial derivative of the m-th component of v; for details, see Section B.2.

The key idea is that the learning problem can be *vectorised* by constructing an objective over the d coordinate directions. Specifically, we define the vectorised surrogate risk as

$$R_t^{\mathrm{vec}}(v) := \sum_{m=1}^d \mathbb{E}_{X_\star, \varepsilon}\left[\Big(S_{\vec{\pi}_{t|\star}}^m v^m(\vec{X}_t, X_\star) - \Psi_t{}^m(\vec{X}_t, X_\star)\Big)^2\right]. \tag{19}$$

Note that the square and sum have been swapped with respect to R_t^c. Thus, while R_t^c penalises the sum of the coordinate residuals, the vectorised objective penalises each coordinate residual separately. Identifying $R_t^{\mathrm{vec}}(v) := \sum_{m=1}^d R_t^{\mathrm{vec},\,m}(v^m)$, minimising $R_t^{\mathrm{vec}}(v)$ is equivalent to minimising, independently for each coordinate $m = 1, \dots, d$, the scalar regression objective $R_t^{\mathrm{vec},\,m}(v^m)$. By Cauchy–Schwarz and (13), $R_t(v) \leq R_t^c(v) \leq d\, R_t^{\mathrm{vec}}(v)$, so the vectorised objective also controls the marginal Stein risk.

In practice, for each coordinate $m = 1, \dots, d$, we consider the regularised empirical risk minimisation problem in the RKHS $\mathcal{H}$:

$$\frac{1}{n}\sum_{j=1}^{n}\left(S^m_{\vec{\pi}_{t|\star}} v^m(\vec{X}_t^{(j)}, X_\star^{(j)}) - \Psi_t{}^m(\vec{X}_t^{(j)}, X_\star^{(j)})\right)^2 + \lambda\|v^m\|^2_{\mathcal{H}}, \tag{20}$$

which again admits a closed-form solution by the representer theorem; the explicit expression is deferred to Proposition B.1 in Section B.2.

This vectorised trick yields alternative Stein transport algorithms. For instance, we present the vectorised counterpart of Stein transport (Algorithm 1) in Section C.1.

5 Theoretical guarantees for Stein transport

In this section, we provide theoretical insights about Stein transport and discuss its connection with (kernelised) score matching.

Score matching. It is instructive to compare the Stein transport objective in Proposition 4.1 with denoising score matching, the central loss used in score-based generative modelling (Song et al., 2021). We state the comparison along the forward OU path $(\vec{\pi}_t)_{t\in[0,T]}$, where the conditional noising distributions are explicit. By the denoising score matching identity (Vincent, 2011), minimising $\|s - \nabla \log \vec{\pi}_t\|^2_{L^2(\vec{\pi}_t)}$ is equivalent, up to an additive constant independent of s, to minimising

$$\mathcal{L}_c^{SM}(s) := \mathbb{E}_{X_\star,\varepsilon_t}\left[\left\|s(\vec{X}_t) + \frac{\varepsilon_t}{\sigma_t}\right\|^2\right], \tag{21}$$

where $\vec{X}_t = e^{-t}X_\star + \sigma_t\varepsilon_t$. The score-matching objective and its conditional counterpart have the same minimisers, since they differ only by a constant independent of s; see Appendix D.5. This equivalence does not generally hold for the Stein transport objective and its conditional counterpart.

In an RKHS, the regularised empirical problem decouples over coordinates:

$$\frac{1}{n}\sum_{j=1}^{n}\left(s^m(\vec{X}_t^{(j)}) + \frac{\varepsilon_t^{m,(j)}}{\sigma_t}\right)^2 + \lambda\|s^m\|^2_{\mathcal{H}}, \qquad m = 1, \dots, d, \tag{22}$$

which is standard kernel ridge regression (Kanagawa et al., 2018, Section 3.2). Its minimiser is

$$s_t^{m,*} = \frac{1}{n}\sum_{j=1}^{n}\phi^{m,j}k(\cdot, \vec{X}_t^{(j)}), \qquad \left(\frac{1}{n}K + \lambda I_n\right)\phi^m = -\frac{\varepsilon_t^m}{\sigma_t}, \tag{23}$$

where $K_{ij} = k(\vec{X}_t^{(i)}, \vec{X}_t^{(j)})$. The resulting score estimate can be used in reverse-time stochastic or deterministic dynamics; we refer to these baselines as kernel score matching and kernel probability flow, respectively, and give details in Section C.2.

Bias-variance tradeoff. Both Stein transport and score matching lead to regularised least-squares problems, but with different regression targets. Kernelised denoising score matching regresses against the conditional space score $-\frac{\varepsilon_t}{\sigma_t}$, using $\vec{X}_t$ as input. By contrast, Stein transport regresses the conditional Stein operator against the target Ψ_t defined above, with the regression design depending on the noised sample $\vec{X}_t$ and the conditioning variable $X_\star$. This difference affects the noise level of the corresponding least-squares problems. In particular, the score matching target remains random even after conditioning on $\vec{X}_t$, whereas the Stein target is deterministic once the variables entering the conditional Stein regression are fixed:

Proposition 5.1. ***(a)*** *The score matching target is generally noisy:* $\mathrm{Cov}\left(-\frac{\varepsilon_t}{\sigma_t}\,\middle|\,\vec{X}_t\right) > 0$*, except in degenerate cases. Moreover, if* $X_\star \sim \mathcal{N}(0, I)$*, then this conditional covariance is* $\mathcal{O}(t^{-1})$ *as* $t \to 0$*.*

(b) *The Stein target* Ψ_t *is a deterministic function of* $(X_\star, \varepsilon_t)$*, and therefore* $\mathrm{Var}(\Psi_t \mid X_\star, \varepsilon_t) = 0$*.*

The proof is deferred to Section D.1.6 and relies on rewriting Stein transport in ridge regression form and comparing it with score matching. This comparison highlights a possible bias-variance tradeoff: the conditional Stein surrogate may introduce bias, while its regression target is deterministic given the augmented input. This distinction may contribute to the empirical robustness observed below, although it does not by itself establish a lower variance for the fitted estimator.

We now analyse the excess conditional Stein risk at a fixed time $t > 0$; this does not directly provide an error bound for the final generated distribution.

In the following, the domain of $S_{\pi_{t|\star}}$ is the Sobolev space $W^{1,2}(\vec{\pi}_t) = \{v \in L^2(\vec{\pi}_t) : \nabla v \in L^2(\vec{\pi}_t)\}$; see (24) and (25) in the appendix for precise definitions. We now state the assumptions required for the proof of Theorem 5.6.

Assumption 5.2. $\mathbb{E}[\|X_\star\|^2] < \infty$.

Assumption 5.3 (Kernel regularity)**.** The kernel k is twice continuously differentiable, and there exists $M > 0$ such that k and all its first- and second-order partial derivatives are bounded by M.

Assumption 5.3 is satisfied by many commonly used kernels, including the Gaussian (RBF) kernel.

Assumption 5.4. $\mathcal{H}^d$ is dense in $W^{1,2}(\vec{\pi}_t)$.

For every fixed $t > 0$, Assumption 5.4 is satisfied by the Gaussian kernel with any positive bandwidth. Indeed, the Gaussian RKHS is dense in $C_0^1(\mathbb{R}^d)$, with uniform convergence of functions and their first derivatives, by Simon-Gabriel and Schölkopf (2018, Theorem 17). Since $\vec{\pi}_t$ has a smooth, strictly positive density, $C_c^\infty(\mathbb{R}^d)$ is dense in $W^{1,2}(\vec{\pi}_t)$. These two approximation properties imply the required density of $\mathcal{H}^d$ in $W^{1,2}(\vec{\pi}_t)$.

Let $\widetilde{\mathcal{H}}_t = S_{\pi_{t|\star}}\mathcal{H}^d \subset L^2(\pi_{\vec{X}_t,X_\star})$ be the Stein-transformed RKHS (see Proposition D.1). By the orthogonal decomposition $L^2(\pi_{\vec{X}_t,X_\star}) = \overline{\widetilde{\mathcal{H}}_t} \overset{\perp}{\oplus} \widetilde{\mathcal{H}}_t^\perp$, the Stein target admits a unique decomposition $\Psi_t = h_t + h_t^\perp$ where $h_t \in \overline{\widetilde{\mathcal{H}}_t}$ and $h_t^\perp \in \widetilde{\mathcal{H}}_t^\perp$. Here, h_t is the component of the target that can be approximated by conditional Stein features, while $h_t^\perp$ is the residual orthogonal to their closed span. The following assumption concerns the integrability of $h_t^\perp$.

Assumption 5.5. There exists $q > 1$ such that $h_t^\perp \in L^{2q}(\pi_{\vec{X}_t,X_\star})$.

The following theorem bounds the error of the estimator $\hat{v}_{t,\lambda}$ relative to the minimum error on $W^{1,2}(\vec{\pi}_t)$.

Theorem 5.6. *Fix $t > 0$ and $0 < \lambda \le e^{-1}$. Suppose that Assumptions 5.2 to 5.5 are true. Then,*

$$\begin{aligned}
&\mathbb{E}\left[R_t^c(\hat{v}_{t,\lambda}) - \inf_{v \in W^{1,2}(\vec{\pi}_t)} R_t^c(v)\right] \\
&\lesssim \left(\left(1 + \frac{\log(1/\lambda)}{\lambda n}\right)^2 + \lambda^{1/(q-1)}\right) \inf_{g \in \widetilde{\mathcal{H}}_t} \left\{\|h_t - g\|^2_{L^2(\pi_{\vec{X}_t,X_\star})} + \lambda\|g\|^2_{\widetilde{\mathcal{H}}_t}\right\} \\
&+ \frac{\log(1/\lambda)}{\lambda n}\left(2 + \frac{\log(1/\lambda)}{\lambda n}\right) + \left(\lambda^{1/(q-1)} + \lambda^2\right).
\end{aligned}$$

The expectation is over the training sample, and the implicit constant may depend on t and the fixed problem parameters, but not on n or λ. Extending the analysis to bound the cumulative excess risk from $t = 0$ to $t = T$ is left for future work. For the vectorised objective, under the corresponding assumptions for each directional regression problem, the same argument yields coordinatewise excess-risk bounds. Summing these bounds preserves the convergence rate in n, with a factor d when the coordinatewise constants are uniformly bounded.

Let $T : L^2(\pi_{\vec{X}_t,X_\star}) \to L^2(\pi_{\vec{X}_t,X_\star})$, $u \to \int u(z)\widetilde{k}_t(.,z)\mathrm{d}\pi_{\vec{X}_t,X_\star}(z)$ denote the integral operator in $L^2(\pi_{\vec{X}_t,X_\star})$ associated with $\widetilde{k}_t$. For $r > 0$, we write $[\widetilde{\mathcal{H}}_t]^r = \mathrm{Range}(T^r)$. We refer to Steinwart and Scovel (2012) for the theory on powers of kernels and integral operators.

Corollary 5.7. *Suppose the assumptions of Theorem 5.6 hold. In addition suppose that $h_t \in [\widetilde{\mathcal{H}}_t]^r$ for $0 < r < \frac{1}{2}$. Then, if we choose $\lambda = \frac{1}{n^{1/2}}$,*

$$\mathbb{E}\left[R_t^c(\hat{v}_{t,\lambda}) - \inf_{v \in W^{1,2}(\vec{\pi}_t)} R_t^c(v)\right] = O\left(\frac{1}{n^r}\right).$$

A proof of Corollary 5.7 is given in Section D.4. The rate obtained is comparable to the rate we can obtain for the score estimator in kernel score matching, see Corollary D.16.

6 Experiments

In this section, we evaluate the proposed methods against several competing approaches on both toy and real-data experiments. We compare six generative methods: deterministic Stein transport (DST, Algorithm 1 with $\alpha = 0$), vectorised deterministic Stein transport (VDST, Algorithm 2 with $\alpha = 0$), stochastic Stein transport (SST, Algorithm 1 with $\alpha = 1$), vectorised stochastic Stein transport (VSST, Algorithm 2 with $\alpha = 1$), kernel score matching (SM, Algorithm 3), and kernel probability flow (PF, Algorithm 4). Computational cost and inference latency are discussed in Section E.3.

Synthetic Gaussian mixture. We first consider a four-component Gaussian mixture model (GMM) in dimension $d = 10$. We also include a kernel density estimation (KDE) baseline, which generates samples directly from its fitted Gaussian mixture. We use training set sizes ranging from 40 to 970. For each training size, all methods use the same training dataset, and the performance is evaluated using 2,000 generated samples against a reference dataset of size 2,000. A grid search over (s, λ) is performed for each training size; see Section E.4 for details. Figure 1 shows marginal density

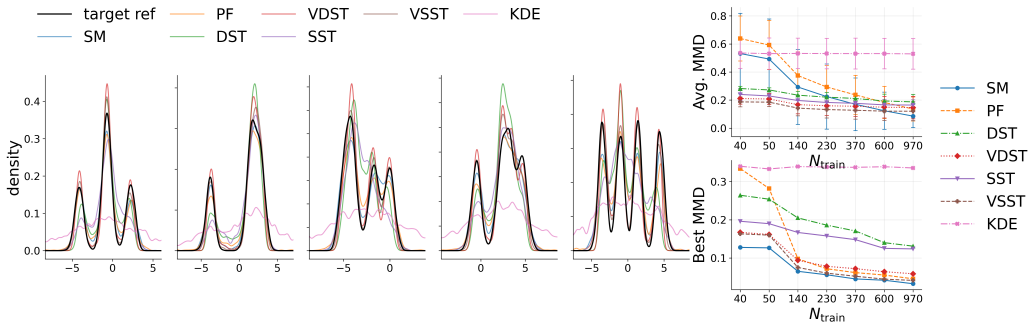


Figure 1: 10D Gaussian mixture. Left: kernel density estimates of the first five marginals. Right: mean MMD (top) and minimum MMD (bottom) versus N_{train}. Statistics are computed over the hyperparameter grid.

estimates at $N_{\text{train}} = 970$ and maximum mean discrepancy (MMD) versus training size. Vectorised variants generally achieve lower grid-averaged MMD than their non-vectorised counterparts, while stochastic variants achieve lower MMD than their deterministic counterparts in this experiment. Although SM achieves the lowest MMD at the best configuration in the tested grid, we found that Stein transport methods were much more robust across the parameter grid, see Section E.4. Additionally, for smaller training sizes (approximately 40 to 230), all Stein transport methods achieve lower parameter-averaged MMD compared to SM and PF.

Posterior emulation. We next learn from a limited set of approximate posterior samples generated by HMC (Betancourt, 2017), aiming to generate additional approximate draws without further MCMC simulation.

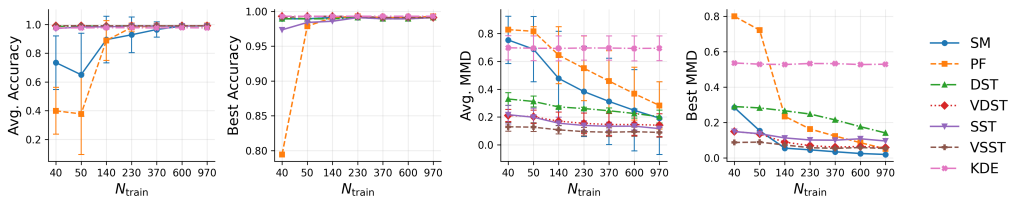


Figure 2: BLR on the breast cancer dataset. From left to right: mean accuracy, maximum accuracy, mean MMD, and minimum MMD versus N_{train}. Statistics are computed over the hyperparameter grid.

We consider Bayesian Logistic Regression (BLR) on the UCI `breast_cancer` dataset ($d = 30$) with Gaussian prior; see Section E.5 for details. Using $2{,}000$ HMC samples as reference, we evaluate VDST, DST, VSST, SST, SM, PF, and KDE across training set sizes ranging from 40 to 970. Methods are assessed using predictive accuracy of the Bayesian model average and MMD relative to the HMC reference samples. Results in Figure 2 are consistent with those of the GMM example, with Stein transport methods exhibiting greater robustness in this more challenging setting and good performance especially in the small data regime.

Generative data augmentation. We finally consider a generative data augmentation (GDA) experiment for real-valued datasets. GDA (Yamaguchi et al., 2020; Zheng et al., 2023) enlarges a training dataset by adding synthetic input–response pairs. Following recent evaluation protocols for synthetic data generation (Kotelnikov et al., 2023; Manousakas and Aydöre, 2023), we evaluate both the fidelity of the generated samples and their utility for downstream learning tasks, referred to as ML (machine learning) efficiency. For conciseness, we focus on VDST and SM as representatives of the two approaches. We also include additional neural-network-based generative modeling baselines (Xu et al., 2019) for tabular data: a generative adversarial network (CTGAN) and a variational autoencoder (TVAE). We work on four tabular regression datasets: Yacht Hydrodynamics (YAC), Diabetes (DIA), California Housing (CAL) and Song Popularity (SON). The reader may refer to Section E.6 for additional details.

	YAC	DIA	CAL	SON
Real	1.05 $\pm$ 0.32	58.69 $\pm$ 3.92	0.41 $\pm$ 0.02	15.23 $\pm$ 0.76
VDST	**3.69** $\pm$ **1.02**	**58.09** $\pm$ **4.54**	**0.45** $\pm$ **0.02**	16.16 $\pm$ 0.79
SM	5.29 $\pm$ 1.91	59.09 $\pm$ 3.10	0.47 $\pm$ 0.04	**16.06** $\pm$ **0.62**
CTGAN	17.98 $\pm$ 2.78	77.90 $\pm$ 9.44	0.84 $\pm$ 0.06	16.85 $\pm$ 1.09
TVAE	8.88 $\pm$ 2.09	65.43 $\pm$ 6.22	0.49 $\pm$ 0.04	17.04 $\pm$ 0.82

Table 1: RMSE obtained by training Random Forest on real versus synthetic data.

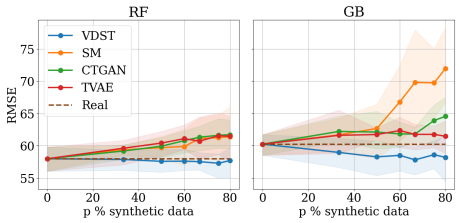


Figure 3: Diabetes RMSE versus synthetic proportion $p = N_{\text{synthetic}}/(N_{\text{real}} + N_{\text{synthetic}})$ (%).

We evaluate downstream utility by training models on synthetic data only and testing on held-out real data, using real training sets of the same size as a baseline. For each of five random $80/20$ train/test splits, we generate five synthetic datasets per method, train a model on each, and report mean RMSE and standard deviation across splits. Results for a Random Forest model are shown in Table 1, while additional models are deferred to Section E.6. Among the synthetic-data methods, VDST achieves the lowest mean RMSE on three of the four datasets.

We next augment a fixed real training set with increasing numbers of synthetic samples. Figure 3 shows RMSE versus the synthetic proportion, with a dashed line for training on real data alone. Shaded regions denote 20% quantiles. Additional datasets and models are deferred to Section E.6. We observe that augmenting real data with synthetic samples does not systematically improve predictive performance, and may even degrade it. Note that these findings are consistent with previous studies on ML efficiency for tabular synthetic data (Manousakas and Aydöre, 2023; Kotelnikov et al., 2023). On Diabetes, VDST augmentation yields lower mean RMSE than the real-data baseline at several augmentation levels for both Random Forest and Gradient Boosting, although the uncertainty bands overlap.

Limitations and outlook. While our focus is on small-sample regimes, extending the proposed methods to larger datasets would benefit from scalable kernel approximations such as Nyström or random-feature methods. Another promising direction is to develop principled hybrid objectives interpolating between score matching and Stein transport, combining the strengths of both approaches. Finally, a complete theoretical analysis should couple the statistical properties of the regression problems with the dynamical error induced by the learned transport, for instance through a mean-field analysis of the generated particle system.

Statement of AI Usage. Version 5.5 of ChatGPT was used to help correct the proof of Theorem 5.6, following the discovery of errors in it after the initial submission. It assisted in generating ideas and carrying out calculations to establish intermediate results in the proof of Theorem 5.6, including the lemmas in Section D.2. However, the idea of distinguishing between the cases of bounded and unbounded kernels in the expectation analysis originated with the authors. The same version of ChatGPT was also used to perform probability and integral calculations in Sections D.1.2 and D.1.3. The same version of Codex available through ChatGPT was used to implement the algorithms and conduct the numerical experiments in Section 6.

We confirm that the overall structure of the proof, the steps required to develop it, and the design and interpretation of the numerical experiments are entirely the work of the authors. The authors take full responsibility for all results presented in this paper.

7 Acknowledgments

A. Korba and C. Chazal acknowledge the support of ANR PEPR PDE-AI.

L. Wang acknowledges the King's Computational Research, Engineering and Technology Environment (CREATE) for the support of computational resources.

References

Albergo, M., Boffi, N. M., and Vanden-Eijnden, E. (2025). Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80.

Anastasiou, A., Barp, A., Briol, F.-X., Ebner, B., Gaunt, R. E., Ghaderinezhad, F., Gorham, J., Gretton, A., Ley, C., Liu, Q., et al. (2023). Stein's method meets computational statistics: A review of some recent developments. *Statistical Science*, 38(1):120–139.

Bach, F. (2024). *Learning theory from first principles*. MIT press.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., and Salim, A. (2023). The probability flow ode is provably fast. *Advances in Neural Information Processing Systems*, 36:68552–68575.

Haussmann, U. G. and Pardoux, E. (1986). Time reversal of diffusions. *The Annals of Probability*, pages 1188–1205.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851.

Huang, Y., Transue, T., Wang, S.-H., Feldman, W. M., Zhang, H., and Wang, B. (2025). Improving flow matching by aligning flow divergence. In *Forty-second International Conference on Machine Learning*.

Hyvärinen, A. and Dayan, P. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4).

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

Kotelnikov, A., Baranchuk, D., Rubachev, I., and Babenko, A. (2023). Tabddpm: Modelling tabular data with diffusion models. In *International conference on machine learning*, pages 17564–17579. PMLR.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. (2023). Flow matching for generative modeling. *International Conference on Learning Representations*.

Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 29.

Liu, X., Gong, C., and Liu, Q. (2023). Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations, ICLR*.

Manousakas, D. and Aydöre, S. (2023). On the usefulness of synthetic tabular data generation. *Data-centric Machine Learning Research (DMLR) Workshop at the 40th International Conference on Machine Learning*.

Nüsken, N. (2024). Stein transport for bayesian inference. *arXiv preprint arXiv:2409.01464*.

Oates, C. J., Girolami, M., and Chopin, N. (2017). Control functionals for monte carlo integration. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):695–718.

Pavliotis, G. A. (2014). Stochastic processes and applications. *Texts in applied mathematics*, 60.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer.

Simon-Gabriel, C.-J. and Schölkopf, B. (2018). Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research*, 19(44):1–29.

Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

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

Steinwart, I. and Scovel, C. (2012). Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35:363–417.

Trabucco, B., Doherty, K., Gurinas, M., and Salakhutdinov, R. (2024). Effective data augmentation with diffusion models. *International Conference on Learning Representations*.

Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wang, Z., Wang, P., Liu, K., Wang, P., Fu, Y., Lu, C.-T., Aggarwal, C. C., Pei, J., and Zhou, Y. (2025). A comprehensive survey on data augmentation. *IEEE Transactions on Knowledge and Data Engineering*.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32.

Yamaguchi, S., Kanai, S., and Eda, T. (2020). Effective data augmentation with multi-domain learning gans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6566–6574.

Yu, H., Klami, A., Hyvärinen, A., Korba, A., and Chehab, O. (2025). Density ratio estimation with conditional probability paths. *International Conference on Machine Learning*.

Zheng, C., Wu, G., and Li, C. (2023). Toward understanding generative data augmentation. *Advances in neural information processing systems*, 36:54046–54060.

A Proof of Proposition 4.1

We use the fact that the conditional distribution of $\vec{X}_t$ given $X_\star$ is $\vec{\pi}_{t|\star} = \mathcal{N}(e^{-t}X_\star, \sigma_t^2 I_d)$ (see (2)). With this, we see that

$$\begin{aligned}
\partial_t \log \vec{\pi}_t(x_t) &= \frac{\partial_t \vec{\pi}_t(x_t)}{\vec{\pi}_t(x_t)} \\
&= \frac{1}{\vec{\pi}_t(x_t)} \partial_t \int \vec{\pi}_{t|\star}(x_t)\vec{\pi}_*(X_\star) dX_\star \\
&= \int (\partial_t \vec{\pi}_{t|\star}(x_t|X_\star)) \frac{\pi_\star(X_\star)}{\vec{\pi}_t(x_t)} dX_\star \\
&= \int (\partial_t \log \vec{\pi}_{t|\star}(x_t|X_\star)) \frac{\pi_\star(X_\star)\vec{\pi}_{t|\star}(x_t|X_\star)}{\vec{\pi}_t(x_t)} dX_\star \\
&= \int (\partial_t \log \vec{\pi}_{t|\star}(x_t|X_\star)) \vec{\pi}_{X_\star|x_t}(X_\star) dX_\star \\
&= \mathbb{E}_{X_\star|x_t}\left[\partial_t \log \vec{\pi}_{t|\star}(x_t)\right].
\end{aligned}$$

and with similar calculations $\frac{\Delta \vec{\pi}_t(x_t)}{\vec{\pi}_t(x_t)} = \mathbb{E}_{X_\star|x_t}\left[\frac{\Delta \vec{\pi}_{t|\star}(x_t|X_\star)}{\vec{\pi}_{t|\star}(x_t|x_\star)}\right]$ so $-\partial_t \log \vec{\pi}_t(x_t) - \alpha \frac{\Delta \vec{\pi}_t}{\vec{\pi}_t} = \mathbb{E}_{X_\star|x_t}[\Psi_t(x_t, X_\star)]$. We also have (as standardly used in SM), $\nabla \log \vec{\pi}_t(x_t) = \mathbb{E}_{X_\star|x_t}\left[\nabla \log \vec{\pi}_{t|\star}(x_t|X_\star)\right]$. Hence,

$$\begin{aligned}
S_{\vec{\pi}_t} v\,(x_t) &= \nabla \log \vec{\pi}_t(x_t) \cdot v(x_t) + \nabla \cdot v(x_t) \\
&= \mathbb{E}_{X_\star|x_t}\left[\nabla \log \vec{\pi}_{t|\star}(x_t|X_\star)\right] \cdot v(x_t) + \nabla \cdot v(x_t) \\
&= \mathbb{E}_{X_\star|x_t}\left[\nabla \log \vec{\pi}_{t|\star}(x_t|X_\star) \cdot v(x_t) + \nabla \cdot v(x_t)\right] \\
&= \mathbb{E}_{X_\star|x_t}\left[(S_{\vec{\pi}_{t|\star}} v)(x_t|X_\star)\right].
\end{aligned}$$

Therefore, by Jensen's inequality, the objective can be written and upper-bounded as

$$\begin{aligned}
&\mathbb{E}_{\vec{X}_t \sim \vec{\pi}_t}\left[\left(S_{\vec{\pi}_t} v\,(\vec{X}_t) + \partial_t \log \vec{\pi}_t(\vec{X}_t) + \alpha \frac{\Delta \vec{\pi}_t(\vec{X}_t)}{\vec{\pi}_t(\vec{X}_t)}\right)^2\right] \\
&= \mathbb{E}_{\vec{X}_t \sim \vec{\pi}_t}\left[\left(\mathbb{E}_{X_\star|\vec{X}_t}\left[S_{\vec{\pi}_{t|\star}} v\,(\vec{X}_t|X_\star) - \Psi_t(\vec{X}_t, X_\star)\right]\right)^2\right] \\
&\leq \mathbb{E}_{\vec{X}_t \sim \vec{\pi}_t}\left[\mathbb{E}_{X_\star|X_t}\left[(S_{\vec{\pi}_{t|\star}} v\,(X_t|X_\star) - \Psi_t(\vec{X}_t, X_\star))^2\right]\right] \\
&= \mathbb{E}_{X_\star, \varepsilon_t}\left[\left(S_{\vec{\pi}_{t|\star}} v\,(X_\star e^{-t} + \sigma_t \varepsilon_t|X_\star) - \Psi_t(X_\star e^{-t} + \sigma_t \varepsilon, X_\star)\right)^2\right].
\end{aligned}$$

The identities (11) and (12) follow by direct computation from the Gaussian kernel in (2).

B Additional details for Section 4

B.1 Solution in an RKHS for Stein transport

We provide here the proof for Proposition 4.2. Applying the representer theorem (Theorem 3 in Nüsken (2024)) to $u_j'(v) = S_{\vec{\pi}_{t|\star}} v(X_\star^{(j)} e^{-t} + \sigma_t \varepsilon_t^{(j)}|X_\star^{(j)})$ and $y_j = \Psi_t(X_\star^{(j)} e^{-t} + \sigma_t \varepsilon_t^{(j)}, X_\star^{(j)})$ we obtain a solution to (16) which is, following the proof of Appendix A.1.1 in Nüsken (2024):

$$\hat{v}_{t,\lambda} = \frac{1}{n} \sum_{j=1}^n \phi_j \left(k(., \vec{X}_t^{(j)}) \nabla \log \vec{\pi}_{t|\star}(\vec{X}_t^{(j)}|X_\star^{(j)}) + \nabla_2 k(., \vec{X}_t^{(j)})\right)$$

$$\text{s.t} \quad (\frac{1}{n}\Xi_t + \lambda I_n)\phi = \mathbf{\Psi}_t.$$

The previous expression can be simplified using that

$$\nabla \log \vec{\pi}_{t|\star}(\vec{X}_t^{(j)}|X_\star^{(j)}) = -\frac{\vec{X}_t^{(j)} - e^{-t}X_\star^{(j)}}{\sigma_t^2} = \frac{-\varepsilon_t^{(j)}}{\sigma_t}.$$

B.2 Solution for vectorised Stein transport

Let $X_\star$ be a sample from the data distribution. We now detail the different objects appearing in the surrogate objective. A noised sample can be written as:

$$\vec{X}_t := X_\star e^{-t} + \sigma_t \varepsilon_t$$

The conditional Stein operator (evaluated at a noised sample) can be written as:

$$S_{\vec{\pi}_{t|\star}} v(\vec{X}_t, X_\star) = (\nabla \cdot v)(\vec{X}_t) - \frac{\varepsilon_t}{\sigma_t} \cdot v(\vec{X}_t).$$

The directional conditional Stein operator (evaluated at a noised sample) is:

$$S^m_{\vec{\pi}_{t|\star}} v(\vec{X}_t|X_\star) = (\partial_m v^m)(\vec{X}_t) - \frac{\varepsilon_t^m}{\sigma_t} v^m(\vec{X}_t), \qquad m = 1, \dots, d.$$

We have

$$\sum_{m=1}^d S^m_{\vec{\pi}_{t|\star}} v^m(\vec{X}_t|X_\star) = S_{\vec{\pi}_{t|\star}} v(\vec{X}_t|X_\star).$$

The Stein target, evaluated at the noised sample, is

$$\begin{aligned}\Psi_t(\vec{X}_t, X_\star) &= -\partial_t \log \vec{\pi}_{t|\star}(\vec{X}_t|X_\star) - \alpha \frac{\Delta \vec{\pi}_{t|\star}(\vec{X}_t|X_\star)}{\vec{\pi}_{t|\star}(\vec{X}_t|X_\star)} \\ &= \frac{e^{-t}}{\sigma_t}\langle X_\star, \varepsilon_t\rangle + \frac{e^{-2t}+\alpha}{\sigma_t^2}\big(d - \|\varepsilon_t\|^2\big).\end{aligned}$$

We also introduce the directional term:

$$\Psi_t{}^m(\vec{X}_t, X_\star) = \frac{e^{-t}}{\sigma_t} X_\star^m \varepsilon_t^m + \frac{e^{-2t}+\alpha}{\sigma_t^2}\big(1 - (\varepsilon_t^m)^2\big), \qquad m = 1, \dots, d.$$

so that

$$\sum_{m=1}^d \Psi_t{}^m(\vec{X}_t, X_\star) = \Psi_t(\vec{X}_t, X_\star)$$

Hence the surrogate Stein objective can be written as:

$$\mathbb{E}_{X_\star, \varepsilon_t}\left[\Big(S_{\pi_{t|\star}} v(\vec{X}_t|X_\star) - \Psi_t(\vec{X}_t, X_\star)\Big)^2\right] = \mathbb{E}_{X_\star, \varepsilon_t}\left[\left(\sum_{m=1}^d \Big(S^m_{\pi_{t|\star}} v^m(\vec{X}_t|X_\star) - \Psi_t{}^m(\vec{X}_t, X_\star)\Big)\right)^2\right]$$

and the vectorised Stein objective can be written as:

$$\sum_{m=1}^d \mathbb{E}_{X_\star, \varepsilon_t}\left[\Big(S^m_{\pi_{t|\star}} v^m(\vec{X}_t|X_\star) - \Psi_t^m(\vec{X}_t, X_\star)\Big)^2\right].$$

Below, we provide the closed-form minimiser for the empirical regularised version of the vectorised Stein objective.

Proposition B.1. *The unique minimiser of* (20) *in the RKHS $\mathcal{H}$ is given by*

$$\hat{v}^m_{t,\lambda} = \frac{1}{n}\sum_{j=1}^n \phi^{m,j}\left(-k(\cdot, \vec{X}_t^{(j)})\frac{\varepsilon_t^{m,(j)}}{\sigma_t} + \frac{\partial k}{\partial_2 x^m}(\cdot, \vec{X}_t^{(j)})\right),$$

$$\textit{s.t} \qquad \left(\frac{1}{n}\Xi_t^m + \lambda I_n\right)\phi^m = \mathbf{\Psi}_t^m,$$

where $\mathbf{\Psi}_t^m = \big(\Psi_t^m(\vec{X}_t^{(j)}, X_\star^{(j)})\big)_{j=1}^n$. *Writing* $x_i = \vec{X}_t^{(i)}$ *and* $\varepsilon_i = \varepsilon_t^{(i)}$, *the directional Gram matrix has entries*

$$(\Xi_t^m)_{ij} = \frac{\varepsilon_i^m \varepsilon_j^m}{\sigma_t^2} k(x_i, x_j) + \partial_{1m}\partial_{2m} k(x_i, x_j) - \frac{\varepsilon_i^m}{\sigma_t} \partial_{2m} k(x_i, x_j) - \frac{\varepsilon_j^m}{\sigma_t} \partial_{1m} k(x_i, x_j).$$

Here, ∂_{1m} *and* ∂_{2m} *differentiate the* m*-th coordinate of the first and second kernel arguments, respectively. The full Gram matrix in Proposition 4.2 satisfies* $\Xi_t = \sum_{m=1}^d \Xi_t^m$*; each coordinate problem uses its own* Ξ_t^m.

C Additional details on the algorithms

C.1 Vectorised Stein Transport (VST)

The VST algorithm is as follows.

Algorithm 2 Vectorised Stein Transport (VST)

Require: time grid $\{0 = t_0 < t_1 < \cdots < t_m = T\}$, training samples $(X_\star^{(j)})_{j=1}^n$, number of generated samples p, regularisation parameter $\lambda > 0$, stochastic parameter $\alpha \geq 0$.

Ensure: synthetic samples $(\overleftarrow{X}_T^{(r)})_{r=1}^p$.

1: Initialize $\overleftarrow{X}_0^{(r)} \sim \gamma_d$ for $r = 1, \ldots, p$.
2: **for** $i = 0$ to $m - 1$ **do**
3: Simulate noised training samples $\vec{X}_{T-t_i}^{(j)} = X_\star^{(j)} e^{-(T-t_i)} + \sigma_{T-t_i} \varepsilon_{T-t_i}^{(j)}, \quad \varepsilon_{T-t_i}^{(j)} \sim \gamma_d$, for $j = 1, \ldots, n$.
4: For each coordinate $\ell = 1, \ldots, d$, solve the system with $\Xi_{T-t_i}^\ell$ and $\mathbf{\Psi}_{T-t_i}^\ell$ in Proposition B.1; combine the scalar fields into $\hat{v}_{T-t_i,\lambda}$.
5: Evolve synthetic samples: $\overleftarrow{X}_{t_{i+1}}^{(r)} = \overleftarrow{X}_{t_i}^{(r)} - (t_{i+1} - t_i)\hat{v}_{T-t_i,\lambda}(\overleftarrow{X}_{t_i}^{(r)}) + \sqrt{2\alpha(t_{i+1} - t_i)}\eta_{t_i}^{(r)}$, $\eta_{t_i}^{(r)} \sim \gamma_d$.
6: **end for**

C.2 Algorithms: SM and PF

We provide the algorithms for SM and PF as follows.

Algorithm 3 Kernel Score Matching (SM)

Require: time grid $\{0 = t_0 < t_1 < \cdots < t_m = T\}$, training samples $(X_\star^{(j)})_{j=1}^n$, number of generated samples p, regularisation parameter $\lambda > 0$.

Ensure: synthetic samples $(\overleftarrow{X}_T^{(r)})_{r=1}^p$.

1: Initialize $\overleftarrow{X}_0^{(r)} \sim \gamma_d$ for $r = 1, \ldots, p$.
2: **for** $i = 0$ to $m - 1$ **do**
3: Simulate $\vec{X}_{T-t_i}^{(j)} = X_\star^{(j)} e^{-(T-t_i)} + \sigma_{T-t_i} \varepsilon_{T-t_i}^{(j)}, \quad \varepsilon_{T-t_i}^{(j)} \sim \gamma_d$, for $j = 1, \ldots, n$.
4: Solve the linear system in (23) at forward time $T - t_i$, return the velocity field $\hat{v}_{T-t_i,\lambda}$.
5: Evolve synthetic samples: $\overleftarrow{X}_{t_{i+1}}^{(r)} = \overleftarrow{X}_{t_i}^{(r)} + (t_{i+1} - t_i)(\overleftarrow{X}_{t_i}^{(r)} + 2\hat{v}_{T-t_i,\lambda}(\overleftarrow{X}_{t_i}^{(r)})) + \sqrt{2(t_{i+1} - t_i)}\eta_{t_i}^{(r)}$, $\eta_{t_i}^{(r)} \sim \gamma_d$.
6: **end for**

Algorithm 4 Kernel Probability Flow (PF)

Require: time grid $\{0 = t_0 < t_1 < \cdots < t_m = T\}$, training samples $(X_\star^{(j)})_{j=1}^n$, number of generated samples p, regularisation parameter $\lambda > 0$.

Ensure: synthetic samples $(\overleftarrow{X}_T^{(r)})_{r=1}^p$.

1: Initialize $\overleftarrow{X}_0^{(r)} \sim \gamma_d$ for $r = 1, \dots, p$.
2: **for** $i = 0$ to $m - 1$ **do**
3: Simulate $\vec{X}_{T-t_i}^{(j)} = X_\star^{(j)} e^{-(T-t_i)} + \sigma_{T-t_i} \varepsilon_{T-t_i}^{(j)}, \quad \varepsilon_{T-t_i}^{(j)} \sim \gamma_d$, for $j = 1, \dots, n$.
4: Solve the linear system in (23) at forward time $T - t_i$, return the velocity field $\hat{v}_{T-t_i,\lambda}$.
5: Evolve synthetic samples: $\overleftarrow{X}_{t_{i+1}}^{(r)} = \overleftarrow{X}_{t_i}^{(r)} + (t_{i+1} - t_i)(\overleftarrow{X}_{t_i}^{(r)} + \hat{v}_{T-t_i,\lambda}(\overleftarrow{X}_{t_i}^{(r)}))$.
6: **end for**

D Proof of Theorem 5.6

This section is devoted to the proofs of Theorem 5.6 and Corollary 5.7. It consists of five subsections. The first is a collection of small results useful for proving Theorem 5.6. The second contains two lemmas that provide bounds for classical ridge regression but are adapted to the case where the kernel is unbounded. The third is the proof of Theorem 5.6. The fourth section proves Corollary 5.7. The fifth section provides a comparison with the case of score matching.

Throughout this section, we will use the pairs of random variables $(\vec{X}_t, X_\star)$, $(\vec{X}_t, \varepsilon_t)$, and $(X_\star, \varepsilon_t)$ interchangeably, choosing the one that best suits the situation, given that all three pairs provide the same information.

D.1 Auxiliary results

To simplify notation, define the random variable $z_t = (\vec{X}_t, X_\star)$ with law $\pi_{z_t} = \pi_{\vec{X}_t, X_\star}$ and recall that $\Psi_t(z_t) = -\partial_t \log \vec{\pi}_{t|\star}(\vec{X}_t|X_\star) - \alpha \frac{\Delta \vec{\pi}_{t|\star}(\vec{X}_t|X_\star)}{\vec{\pi}_{t|\star}(\vec{X}_t|X_\star)}$.

Recall that the domain of $S_{\vec{\pi}_{t|\star}}$ is

$$W^{1,2}(\vec{\pi}_t) = \{v \in L^2(\vec{\pi}_t) : \nabla v \in L^2(\vec{\pi}_t)\}, \tag{24}$$

equipped with the norm

$$\|v\|_{W^{1,2}(\vec{\pi}_t)} = \left(\|v\|^2_{L^2(\vec{\pi}_t)} + \sum_{i=1}^d \|\nabla v_i\|^2_{L^2(\vec{\pi}_t)} \right)^{\frac{1}{2}}. \tag{25}$$

The following lemmas are necessary intermediate results in the proof for Theorem 5.6.

D.1.1 Conditional Stein RKHS and kernel

The following proposition states that the space induced by the action of $S_{\vec{\pi}_{t|\star}}$ on $\mathcal{H}^d$ is an RKHS not only over the variable $\vec{X}_t$ but also over the pair $z_t = (\vec{X}_t, X_\star)$. The corresponding kernel is very similar to the Stein kernel in $\vec{X}_t$, but it also takes into account the variable $X_\star$ in which it is linear.

Proposition D.1. *Suppose Assumption 5.3 holds. The space $\widetilde{\mathcal{H}}_t = S_{\vec{\pi}_{t|\star}} \mathcal{H}^d = \{S_{\vec{\pi}_{t|\star}} v,\ v \in \mathcal{H}^d\}$ is an RKHS on $\mathbb{R}^d \times \mathbb{R}^d$ with kernel*

$$\widetilde{k}_t(z_t, z_t') = \langle r_{z_t}, r_{z_t'} \rangle_{\mathcal{H}^d}, \quad r_{z_t} = \nabla \log \vec{\pi}_{t|\star}(\vec{X}_t|X_\star) k(., \vec{X}_t) + \nabla_2 k(., \vec{X}_t).$$

Moreover, $\forall v \in \mathcal{H}^d$ and $z_t = (\vec{X}_t, X_\star) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$S_{\vec{\pi}_{t|\star}} v(\vec{X}_t, X_\star) = \langle v, r_{z_t} \rangle_{\mathcal{H}^d}.$$

Proof. This result can be shown by following the same outline of the proof of Theorem 1 in Oates et al. (2017) for the RKHS with Stein kernel. By (Steinwart and Christmann, 2008, Lemma 4.34 and

Corollary 4.36), Assumption 5.3 implies that every $f \in \mathcal{H}$ is continuously differentiable and, for every $x \in \mathbb{R}^d$ and coordinate m,

$$\partial_{2m} k(\cdot, x) \in \mathcal{H}, \qquad \partial_m f(x) = \langle f, \partial_{2m} k(\cdot, x) \rangle_{\mathcal{H}}.$$

Consequently, $\nabla_2 k(\cdot, x) \in \mathcal{H}^d$. The conditional Gaussian density $\vec{\pi}_{t|\star}$ is smooth and strictly positive for $t > 0$, so $\nabla \log \vec{\pi}_{t|\star}(\vec{X}_t | X_\star)$ is well defined. Thus $\widetilde{\mathcal{H}}_t$ is well defined and $r_{z_t} \in \mathcal{H}^d$.

For $v \in \mathcal{H}^d$, $z_t = (\vec{X}_t, X_\star) \in \mathbb{R}^d \times \mathbb{R}^d$

$$\begin{aligned} S_{\vec{\pi}_{t|\star}} v(\vec{X}_t, X_\star) &= \langle v(\vec{X}_t), \nabla \log_{\vec{\pi}_{t|\star}}(\vec{X}_t | X_\star) \rangle + \nabla \cdot v(\vec{X}_t) \\ &= \langle v, \nabla \log_{\vec{\pi}_{t|\star}}(\vec{X}_t | X_\star) k(., \vec{X}_t) \rangle_{\mathcal{H}^d} + \langle v, \nabla_2 k(., \vec{X}_t) \rangle_{\mathcal{H}^d} \\ &= \langle v, r_{z_t} \rangle_{\mathcal{H}^d}. \end{aligned}$$

In this three-line calculation, the second equality uses the ordinary and derivative reproducing properties coordinatewise, and the final equality follows from the definition of r_{z_t}. In particular,

$$\nabla \cdot v(\vec{X}_t) = \sum_{i=1}^{d} \langle v_i, \partial_{2i} k(\cdot, \vec{X}_t) \rangle_{\mathcal{H}} = \langle v, \nabla_2 k(\cdot, \vec{X}_t) \rangle_{\mathcal{H}^d}.$$

Hence,

$$\widetilde{\mathcal{H}}_t = S_{\vec{\pi}_{t|\star}} \mathcal{H}^d = \left\{ g \in L^2(\pi_{z_t}) \text{ such that } g(z_t) = \langle v, r_{z_t} \rangle_{\mathcal{H}^d},\ v \in \mathcal{H}^d \right\}.$$

Now let us endow $\widetilde{\mathcal{H}}_t$ with a Hilbert space structure. For $g \in \widetilde{\mathcal{H}}_t$, let

$$\|g\|_{\widetilde{\mathcal{H}}_t} = \inf_{v \in \mathcal{H}^d} \{ \|v\|_{\mathcal{H}}, \quad \text{such that } S_{\vec{\pi}_{t|\star}} v = g \}.$$

We conclude using Theorem 4.21 in Steinwart and Christmann (2008), which ensures that with this structure, $\widetilde{\mathcal{H}}_t$ is an RKHS with kernel

$$\widetilde{k}_t(z_t, z_t') = \langle r_{z_t}, r_{z_t'} \rangle_{\mathcal{H}^d}.$$ □

D.1.2 The conditional Stein kernel $\widetilde{k}_t$ is bounded in expectation and with high probability

Lemma D.2. *Let U be a χ^2 random variable with d degrees of freedom. For $l \geq 2d$,*

$$\mathbb{P}(U > l) \leq e^{-\frac{(l-2d)}{3}}.$$

Proof. From Laurent and Massart (2000), Lemma 1,

$$\mathbb{P}\Big(U \geq d + 2\sqrt{dx} + 2x\Big) \leq e^{-x}.$$

If we choose $x = \frac{(l-2d)}{3}$, then $d + 2\sqrt{dx} + 2x \leq l$ so

$$\mathbb{P}(U > l) \leq \mathbb{P}(U > d + 2\sqrt{dx} + 2x) \leq e^{-\frac{(l-2d)}{3}}.$$

□

Lemma D.3. *Fix $t > 0$ and suppose Assumption 5.3 holds. Then*

$$\begin{aligned} \mathbb{E}[\widetilde{k}_t(z_t, z_t)] &\leq 2dM \left(1 + \frac{d}{\sigma_t^2}\right), \\ \mathbb{E}[\widetilde{k}_t(z_t, z_t)^2] &\leq 8d^2M^2 \left(1 + \frac{d(d+2)}{\sigma_t^4}\right). \end{aligned}$$

Moreover, for every $B > 0$,

$$\mathbb{P}\Big(\widetilde{k}_t(z_t, z_t) > B^2\Big) \leq \exp(d) \exp\left[-\frac{\sigma_t^2}{3}\left(\frac{B^2}{2dM} - 1\right)\right].$$

Proof. Write $x = \vec{X}_t$ and

$$b(z_t) = \nabla \log \vec{\pi}_{t|\star}(\vec{X}_t \mid X_\star) = \frac{e^{-t}X_\star - \vec{X}_t}{\sigma_t^2} = -\frac{\varepsilon_t}{\sigma_t}, \qquad \varepsilon_t \sim \mathcal{N}(0, I_d).$$

By Proposition D.1, $r_{z_t} = b(z_t)k(\cdot, x) + \nabla_2 k(\cdot, x)$ and $\widetilde{k}_t(z_t, z_t) = \|r_{z_t}\|^2_{\mathcal{H}^d}$. The ordinary and derivative reproducing identities give

$$\|k(\cdot, x)\|^2_{\mathcal{H}} = k(x, x) \le M, \qquad \|\nabla_2 k(\cdot, x)\|^2_{\mathcal{H}^d} = \sum_{j=1}^{d} \partial_{1j}\partial_{2j}k(x, x) \le dM.$$

Thus, using $\|u + v\|^2 \le 2\|u\|^2 + 2\|v\|^2$ in $\mathcal{H}^d$,

$$\begin{aligned}\widetilde{k}_t(z_t, z_t) &\le 2\|b(z_t)k(\cdot, x)\|^2_{\mathcal{H}^d} + 2\|\nabla_2 k(\cdot, x)\|^2_{\mathcal{H}^d} \\ &\le 2M\big(\|b(z_t)\|^2 + d\big) \\ &\le 2dM\left(1 + \frac{\|\varepsilon_t\|^2}{\sigma_t^2}\right).\end{aligned}$$

This bound does not assume that $\nabla_2 k(x, x)$ vanishes. Set $U = \|\varepsilon_t\|^2 \sim \chi^2_d$. Since $\mathbb{E}U = d$ and $\mathbb{E}U^2 = d(d+2)$, the first moment bound follows immediately. For the second, use $(1+a)^2 \le 2(1+a^2)$ for $a \ge 0$ to obtain

$$\mathbb{E}[\widetilde{k}_t(z_t, z_t)^2] \le 8d^2M^2\left(1 + \frac{\mathbb{E}U^2}{\sigma_t^4}\right).$$

For the tail, put $\ell_B = \sigma_t^2\big(B^2/(2dM) - 1\big)$. The pointwise bound implies $\mathbb{P}(\widetilde{k}_t(z_t, z_t) > B^2) \le \mathbb{P}(U > \ell_B)$. If $\ell_B \ge 2d$, Lemma D.2 gives

$$\mathbb{P}(U > \ell_B) \le \exp\left(-\frac{\ell_B - 2d}{3}\right) \le \exp(d)\exp(-\ell_B/3).$$

If $\ell_B < 2d$, the last expression exceeds 1, so it still bounds $\mathbb{P}(U > \ell_B)$. This proves the stated bound for every $B > 0$. □

D.1.3 Conditional Stein operator is continuous

Lemma D.4. *If ν is a standard Gaussian measure on $\mathbb{R}^d$ and $f \in W^{1,2}(\nu)$, there exists $C > 0$ independent of f such that $\|x \cdot f\|^2_{L^2(\nu)} \le C\left(\|f\|^2_L(\nu)^2 + \int \|Jf(x)\|^2_F \,\mathrm{d}\nu(x)\right)$ where $(x \cdot f) : x \to x^T f(x)$ and Jf is the Jacobian matrix of f.*

Proof. By Cauchy-Schwarz and integration by parts,

$$\begin{aligned}\|x \cdot f\|^2_{L^2(\nu)} &= \int \big(x^T f(x)\big)^2 \,\mathrm{d}\nu(x) \\ &\le \int \|x\|^2\|f(x)\|^2 \mathrm{d}\nu(x) \\ &= \frac{1}{(2\pi)^{d/2}} \int \left(xe^{-\|x\|^2/2}\right)^T \left(x\|f(x)\|^2\right) \mathrm{d}x \\ &= \int \nabla \cdot \left(x\|f(x)\|^2\right) \mathrm{d}\nu(x).\end{aligned}$$

The divergence term is

$$\nabla \cdot \left(x\|f(x)\|^2\right) = d\|f(x)\|^2 + 2x^T Jf(x)f(x),$$

where Jf is the Jacobian matrix of f. Integrating the first term gives

$$\int d\|f(x)\|^2 \mathrm{d}\nu(x) = d\|f\|^2_{L^2(\nu)},$$

The second term satisfies $x^T Jf(x)f(x) \leq \|x\|\|Jf(x)\|_{op}\|f(x)\| \leq \|Jf(x)\|_F\|x\|\|f(x)\|$ where $\|.\|_{op}$ denotes the operator norm and $\|.\|_F$ denotes the Frobenius norm. Using Young's inequality, $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$ applied to $a = 2\|Jf(x)\|_F$ and $b = \|x\|\|f(x)\|$,

$$\begin{aligned} 2\int x^T Jf(x)f(x)\mathrm{d}\nu(x) &\leq \int 2\|Jf(x)\|_F\|x\|\|f(x)\|\mathrm{d}\nu(x) \\ &\leq 2\int \|Jf(x)\|_F^2 \,\mathrm{d}\nu(x) \,+ \frac{1}{2}\int \|x\|^2\|f(x)\|^2\mathrm{d}\nu(x). \end{aligned}$$

Gathering everything gives

$$\begin{aligned} &\int \|x\|^2\|f(x)\|^2\mathrm{d}\nu(x) \leq d\|f\|^2_{L^2(\nu)} + 2\int \|Jf(x)\|_F^2 \,\mathrm{d}\nu(x) + \frac{1}{2}\int \|x\|^2\|f(x)\|^2\mathrm{d}\nu(x) \\ &\Leftrightarrow \int \|x\|^2\|f(x)\|^2\mathrm{d}\nu(x) \leq 2\left(d\|f\|^2_{L^2(\nu)} + 2\int \|Jf(x)\|_F^2 \,\mathrm{d}\nu(x)\right). \end{aligned}$$

So,

$$\|x \cdot f\|^2_{L^2(\nu)} \leq C\left(\|f\|^2_{L^2(\nu)} + \int \|Jf(x)\|_F^2\mathrm{d}\nu(x)\right),$$

where $C = \max\{2d, 4\}$. □

Proposition D.5. *For every fixed $t > 0$, the conditional Stein operator $S_{\vec{\pi}_{t|\star}} : W^{1,2}(\vec{\pi}_t) \to L^2(\pi_{z_t})$ is continuous.*

Proof. As $S_{\vec{\pi}_{t|\star}}$ is linear, it is enough to prove that $\sup_{v\in W^{1,2}(\vec{\pi}_t), \|v\|_{W^{1,2}(\vec{\pi}_t)}=1} \|S_{\vec{\pi}_{t|\star}}v\|_{L^2(\pi_{z_t})} < +\infty$. Indeed, let $v \in W^{1,2}(\vec{\pi}_t)$.

$$\begin{aligned} \|S_{\vec{\pi}_{t|\star}}v\|^2_{L^2(\pi_{z_t})} &= \int (S_{\vec{\pi}_{t|\star}}v(x_t, x_\star))^2 \,\mathrm{d}\pi_{\vec{X}_t, X_\star}(x_t, x_\star) \\ &= \int \left(\langle \nabla \log \vec{\pi}_{t|\star}(x_t|x_\star), v(x_t)\rangle + \nabla \cdot v(x_t)\right)^2 \,\mathrm{d}\pi_{\vec{X}_t, X_\star}(x_t, x_\star) \\ &\leq 2\int \langle \nabla \log \vec{\pi}_{t|\star}(x_t|x_\star), v(x_t)\rangle^2 \,\mathrm{d}\pi_{\vec{X}_t, X_\star}(x_t, x_\star) + 2\int (\nabla \cdot v(x_t))^2 \,\mathrm{d}\vec{\pi}_t(x_t). \end{aligned}$$

Recalling that $\vec{X}_t = e^{-t}X_\star + \sigma_t\varepsilon_t$, with $\varepsilon_t \sim \pi_\varepsilon = \mathcal{N}(0, I_d)$, and that $X_\star$ and ε_t are independent, we have $\nabla \log \vec{\pi}_{t|\star}(x_t|x_\star) = \frac{-\varepsilon_t}{\sigma_t}$. A change of variables from $(x_t, x_\star) \sim \pi_{\vec{X}_t, X_\star}$ to $(x_\star, \epsilon) \sim \pi_{\star,\varepsilon} = \pi_\star \otimes \pi_\varepsilon$ gives

$$\int \langle \nabla \log \vec{\pi}_{t|\star}(x_t|x_\star), v(x_t)\rangle^2 \,\mathrm{d}\pi_{\vec{X}_t, X_\star}(x_t, x_\star) = \int\int \left\langle \frac{\epsilon}{\sigma_t}, v(e^{-t}x_\star + \sigma_t\epsilon)\right\rangle^2 \,\mathrm{d}\pi_\varepsilon(\epsilon)\mathrm{d}\pi_\star(x_\star).$$

For $\pi_\star$-almost every $x_\star$, set $f_{x_\star}(\epsilon) = v(e^{-t}x_\star + \sigma_t\epsilon) \in W^{1,2}(\pi_\varepsilon)$. By the chain rule, $Jf_{x_\star}(\epsilon) = \sigma_t Jv(e^{-t}x_\star + \sigma_t\epsilon)$. Applying Lemma D.4, with $C_d = \max\{2d, 4\}$, gives

$$\begin{aligned} &\int \left\langle \frac{\epsilon}{\sigma_t}, f_{x_\star}(\epsilon)\right\rangle^2 \mathrm{d}\pi_\varepsilon(\epsilon) \\ &\leq \frac{C_d}{\sigma_t^2}\left(\|f_{x_\star}\|^2_{L^2(\pi_\varepsilon)} + \int \|Jf_{x_\star}(\epsilon)\|_F^2\mathrm{d}\pi_\varepsilon(\epsilon)\right). \end{aligned}$$

Here C_d is independent of $x_\star$ and t. Integrating over $\pi_\star$ and using $\|Jf_{x_\star}(\epsilon)\|_F^2 = \sigma_t^2\|Jv(e^{-t}x_\star + \sigma_t\epsilon)\|_F^2$, we obtain

$$\begin{aligned} &\int \langle \nabla \log \vec{\pi}_{t|\star}(x_t|x_\star), v(x_t)\rangle^2\mathrm{d}\pi_{\vec{X}_t, X_\star}(x_t, x_\star) \\ &\leq \frac{C_d}{\sigma_t^2}\|v\|^2_{L^2(\vec{\pi}_t)} + C_d\int \|Jv(x_t)\|_F^2\mathrm{d}\vec{\pi}_t(x_t). \end{aligned}$$

Combining this with $(\nabla \cdot v)^2 \leq d\|Jv\|_F^2$ yields

$$\begin{aligned}\|S_{\vec{\pi}_{t|\star}} v\|^2_{L^2(\pi_{z_t})} &\leq \frac{2C_d}{\sigma_t^2}\|v\|^2_{L^2(\vec{\pi}_t)} + 2(C_d + d)\int \|Jv(x_t)\|_F^2 \mathrm{d}\vec{\pi}_t(x_t) \\ &\leq C_t'\|v\|^2_{W^{1,2}(\vec{\pi}_t)},\end{aligned}$$

where $C_t' = \max\{2C_d/\sigma_t^2,\, 2(C_d + d)\} < \infty$ for every fixed $t > 0$.

□

D.1.4 The conditional Stein RKHS $\widetilde{\mathcal{H}}_t$ is dense in the image of $S_{\vec{\pi}_{t|\star}}$

Proposition D.6. *Suppose Assumption 5.3 and Assumption 5.4 hold. Then $\widetilde{\mathcal{H}}_t$ is dense in $S_{\vec{\pi}_{t|\star}} W^{1,2}(\vec{\pi}_t)$.*

Proof. Let $g \in S_{\vec{\pi}_{t|\star}} W^{1,2}(\vec{\pi}_t)$. By definition, there exists $v \in W^{1,2}(\vec{\pi}_t)$ such that $g = S_{\vec{\pi}_{t|\star}} v$. As $\mathcal{H}^d$ is dense in $W^{1,2}(\vec{\pi}_t)$, there is a sequence $(v_n)_{n\geq 1} \subseteq \mathcal{H}^d$ which converges to v. For $n \geq 1$, denote $g_n = S_{\vec{\pi}_{t|\star}} v_n \in \widetilde{\mathcal{H}}_t$. By continuity of $S_{\vec{\pi}_{t|\star}}$ established in Proposition D.5, the sequence $(g_n)_{n\geq 1}$ converges to g. Thus, $\widetilde{\mathcal{H}}_t$ is dense in $S_{\vec{\pi}_{t|\star}} W^{1,2}(\vec{\pi}_t)$. □

D.1.5 Orthogonal decomposition of the Stein target (Ψ_t) and resulting properties

In this section we introduce the orthogonal decomposition of Ψ_t in $L^2(\pi_{z_t}) = \overline{\widetilde{\mathcal{H}}_t} \overset{\perp}{\oplus} \widetilde{\mathcal{H}}_t^{\perp}$: $\Psi_t = h_t + h_t^{\perp}$ with $h_t \in \overline{\widetilde{\mathcal{H}}_t}$ and $h_t^{\perp} \in \widetilde{\mathcal{H}}_t^{\perp}$.

Proposition D.7. *Suppose Assumption 5.3 and Assumption 5.4 hold. Then* $\inf_{v\in W^{1,2}(\vec{\pi}_t)} R_t^c(v) = \|h_t^{\perp}\|^2_{L^2(\pi_{z_t})}$.

Proof. We can write

$$\begin{aligned}\inf_{v\in W^{1,2}(\vec{\pi}_t)} R_t^c(v) &= \inf_{v\in W^{1,2}(\vec{\pi}_t)} \|S_{\vec{\pi}_{t|\star}} v - \Psi_t\|^2_{L^2(\pi_{z_t})} \\ &= \inf_{g\in S_{\vec{\pi}_{t|\star}} W^{1,2}(\vec{\pi}_t)} \|g - \Psi_t\|^2_{L^2(\pi_{z_t})} \\ &= \inf_{g\in S_{\vec{\pi}_{t|\star}} W^{1,2}(\vec{\pi}_t)} \|g - (h_t + h_t^{\perp})\|^2_{L^2(\pi_{z_t})}.\end{aligned}$$

By Proposition D.6, $\widetilde{\mathcal{H}}_t$ is dense in $S_{\vec{\pi}_{t|\star}} W^{1,2}(\vec{\pi}_t)$. As $\widetilde{\mathcal{H}}_t$ is included in $S_{\vec{\pi}_{t|\star}} W^{1,2}(\vec{\pi}_t)$, we have in fact $\overline{\widetilde{\mathcal{H}}_t} = \overline{S_{\vec{\pi}_{t|\star}} W^{1,2}(\vec{\pi}_t)}$ and finally

$$\inf_{v\in W^{1,2}(\vec{\pi}_t)} R_t^c(v) = \inf_{g\in \overline{\widetilde{\mathcal{H}}_t}} \|g - (h_t + h_t^{\perp})\|^2_{L^2(\pi_{z_t})} = \|h_t^{\perp}\|^2_{L^2(\pi_{z_t})}.$$

□

Lemma D.8 (Control of the orthogonal component)**.** *Fix $t > 0$ and suppose Assumption 5.3 and Assumption 5.2 hold. Then*

$$\mathbb{E}[h_t^{\perp}(z_t)^2] \leq 2d\frac{(e^{-2t} + \alpha)^2}{\sigma_t^4} + \frac{e^{-2t}}{\sigma_t^2}\mathbb{E}[\|X_\star\|^2].$$

Proof. We first verify that the Stein target is square-integrable. Write

$$a_t = \frac{e^{-2t} + \alpha}{\sigma_t^2}, \qquad \beta_t = \frac{e^{-t}}{\sigma_t}, \qquad U = \|\varepsilon_t\|^2,$$

where $\varepsilon_t \sim \mathcal{N}(0, I_d)$ is independent of $X_\star$. The explicit conditional target is

$$\Psi_t(z_t) = -a_t(U - d) + \beta_t\langle \varepsilon_t, X_\star\rangle.$$

Since $U \sim \chi^2_d$, independence and Gaussian symmetry give

$$\begin{aligned}\mathbb{E}[(U-d)^2] &= 2d,\\ \mathbb{E}[\langle \varepsilon_t, X_\star\rangle^2 \mid X_\star] &= \|X_\star\|^2,\\ \mathbb{E}[(U-d)\langle \varepsilon_t, X_\star\rangle \mid X_\star] &= 0.\end{aligned}$$

The last equality follows because, for each fixed value of $X_\star$, the integrand is odd in ε_t. Expanding the square and conditioning on $X_\star$ therefore yields

$$\begin{aligned}\mathbb{E}[\Psi_t(z_t)^2] &= 2da_t^2 + \beta_t^2\mathbb{E}[\|X_\star\|^2]\\ &= 2d\frac{(e^{-2t}+\alpha)^2}{\sigma_t^4} + \frac{e^{-2t}}{\sigma_t^2}\mathbb{E}[\|X_\star\|^2] < \infty.\end{aligned}$$

Thus $\Psi_t \in L^2(\pi_{z_t})$ by Assumption 5.2.

Now use the orthogonal decomposition $\Psi_t = h_t + h_t^\perp$ in $L^2(\pi_{z_t})$. Here the closure of $\widetilde{\mathcal{H}}_t$ is taken in $L^2(\pi_{z_t})$. Since $h_t \in \overline{\widetilde{\mathcal{H}}_t}$ and $h_t^\perp \in \widetilde{\mathcal{H}}_t^\perp = (\overline{\widetilde{\mathcal{H}}_t})^\perp$, continuity of the inner product gives $\langle h_t, h_t^\perp\rangle_{L^2(\pi_{z_t})} = 0$. Pythagoras' identity implies

$$\mathbb{E}[h_t^\perp(z_t)^2] = \mathbb{E}[\Psi_t(z_t)^2] - \mathbb{E}[h_t(z_t)^2] \le \mathbb{E}[\Psi_t(z_t)^2]. \tag{26}$$

Combining this with the computed second moment proves the claim. This argument does not require $h_t \in \widetilde{\mathcal{H}}_t$ or an exact representation $h_t = S_{\vec{\pi}_{t|\star}} v$. □

D.1.6 Solving Stein transport in $\widetilde{\mathcal{H}}_t$

We prove the result in the non-vectorised case, as the notation is simpler and the derivations are similar for the vectorised case.

Proposition D.9 (Stein transport in noise-augmented coordinates)**.** *Suppose Assumption 5.3 holds. Problem* (16) *is equivalent to*

$$\frac{1}{n}\sum_{j=1}^n \Big(g(z_t^{(j)}) - \Psi_t(z_t^{(j)})\Big)^2 + \lambda\|g\|^2_{\widetilde{\mathcal{H}}_t}, \tag{27}$$

in the sense that the unique minimisers of (16) *and* $\hat{v}_{t,\lambda}$ *are related via* $\hat{g}_\lambda((\vec{X}_t, X_\star)) = S_{\vec{\pi}_{t|\star}}\hat{v}_{t,\lambda}(\vec{X}_t)$.

Proof. By Proposition D.1, $S_{\vec{\pi}_{t|\star}} v(\vec{X}_t) = \langle v, r_{z_t}\rangle_{\mathcal{H}^d}$. The objective (27) can now be written in the form

$$\frac{1}{n}\sum_{j=1}^n \big(\langle v, r_{z_t^{(j)}}\rangle_{\mathcal{H}^d} - \Psi_t(z_t^{(j)})\big)^2 + \lambda\|v\|^2_{\mathcal{H}^d},$$

with unique minimiser

$$\hat{v}_{t,\lambda} = \sum_{j=1}^n c_j\, r_{z_t^{(j)}}(\cdot),$$

for some coefficients $c \in \mathbb{R}^n$. The claim now follows from

$$S_{\vec{\pi}_{t|\star}}(\hat{v}_{t,\lambda})(\vec{X}_t) = \sum_{j=1}^n c_j\,\langle r_{z_t^{(j)}}, r_{z_t}\rangle_{\mathcal{H}} = \sum_{j=1}^n c_j\,\widetilde{k}_t(z_t^{(j)}, z_t),$$

noting that the coefficients c solve

$$(\widetilde{K} + n\lambda I)c = \boldsymbol{\Psi_t}, \qquad \widetilde{K}_{j\ell} = \widetilde{k}_t(z_t^{(j)}, z_t^{(\ell)}), \quad \boldsymbol{\Psi}_{t,j} = \Psi_t(z_t^{(j)}).$$ □

This proposition immediately implies part (b) of Proposition 5.1.

For the vectorised objective, each $v^m : \mathbb{R}^d \to \mathbb{R}$ is scalar-valued but depends on the full input. We use the same augmented variable $z_t = (\vec{X}_t, X_\star)$, equivalently $(X_\star, \varepsilon_t)$ with the full noise vector. For $z = (x, x_\star)$, write $\varepsilon_t(z) = (x - e^{-t}x_\star)/\sigma_t$ and define the directional feature and kernel

$$r^m_{t,z}(\cdot) = \partial_{2m}k(\cdot, x) - \frac{\varepsilon^m_t(z)}{\sigma_t}k(\cdot, x),$$
$$\widetilde{k}^m_t(z, z') = \langle r^m_{t,z}, r^m_{t,z'}\rangle_{\mathcal{H}}.$$

Let $\widetilde{\mathcal{H}}^m_t$ be the RKHS associated with $\widetilde{k}^m_t$. In particular, $(\Xi^m_t)_{ij} = \widetilde{k}^m_t(z^{(i)}_t, z^{(j)}_t)$ for the matrix in Proposition B.1.

Proposition D.10 (Vectorised Stein transport in noise-augmented coordinates)**.** *Suppose Assumption 5.3 holds. For each $m = 1, \dots, d$, problem* (20) *is equivalent to*

$$\frac{1}{n}\sum_{j=1}^{n}\left(g^m(z^{(j)}_t) - \Psi^m_t(z^{(j)}_t)\right)^2 + \lambda\|g^m\|^2_{\widetilde{\mathcal{H}}^m_t}, \tag{28}$$

in the sense that the unique minimisers of (20) *and* (28) *satisfy*

$$(g^m)^*(z_t) = S^m_{\vec{\pi}_{t|\star}}(v^m)^*(\vec{X}_t|X_\star).$$

Proof. The derivative reproducing property gives $S^m_{\vec{\pi}_{t|\star}}f(x|x_\star) = \langle f, r^m_{t,z}\rangle_{\mathcal{H}}$. The feature-map construction equips $\widetilde{\mathcal{H}}^m_t$ with the norm

$$\|g\|_{\widetilde{\mathcal{H}}^m_t} = \inf\{\|f\|_{\mathcal{H}} : g(z) = \langle f, r^m_{t,z}\rangle_{\mathcal{H}} \text{ for all } z\}.$$

Thus the objectives agree when f is the minimum-norm preimage of g, yielding the claimed correspondence. □

From Propositions D.9 and D.10, part (b) of Proposition 5.1 follows: the target Ψ_t is a deterministic function of the augmented input z_t, equivalently of $(X_\star, \varepsilon_t)$, so its conditional variance is zero.

For part (a), using $\vec{X}_t = e^{-t}X_\star + \sigma_t\varepsilon_t$, we obtain

$$\mathrm{Cov}\left(-\frac{\varepsilon_t}{\sigma_t}\,\middle|\,\vec{X}_t\right) = \frac{e^{-2t}}{\sigma^4_t}\,\mathrm{Cov}(X_\star \mid \vec{X}_t).$$

Thus the target generally retains conditional randomness. Here positivity means a nonzero positive semidefinite covariance matrix; strict positive definiteness additionally requires nondegeneracy in every direction. If $X_\star \sim \mathcal{N}(0, I_d)$, the joint Gaussian conditioning formula gives $\mathrm{Cov}(X_\star \mid \vec{X}_t) = \sigma^2_t I_d$, hence

$$\mathrm{Cov}\left(-\frac{\varepsilon_t}{\sigma_t}\,\middle|\,\vec{X}_t\right) = \frac{e^{-2t}}{1 - e^{-2t}}I_d \sim \frac{1}{2t}I_d \qquad \text{as } t \to 0.$$

This proves the stated $\mathcal{O}(t^{-1})$ growth (in operator norm).

D.2 Kernel ridge regression estimates

This section collects the kernel ridge regression estimates that constitute the main technical ingredients in the proof of Theorem 5.6. Throughout this section we consider the kernel ridge regression operator

$$\mathcal{S}_{n,\lambda} : L^2(\pi_{z_t}) \to \widetilde{\mathcal{H}}_t,$$

defined by

$$\mathcal{S}_{n,\lambda}f = (\widehat{\Sigma}_{\widetilde{k}_t} + \lambda I)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(z^{(i)})\widetilde{k}_t(\cdot, z^{(i)})\right).$$

where $\widehat{\Sigma}_{\widetilde{k}_t} = \frac{1}{n}\sum_{i=1}^{n}\widetilde{k}_t(., z^{(i)}) \otimes \widetilde{k}_t(., z^{(i)})$ is the empirical covariance operator. Since the conditional Stein transport problem can be rewritten as a kernel ridge regression problem over the RKHS

$\widetilde{\mathcal{H}}_t$, our analysis follows the standard operator-theoretic approach for KRR. The only additional difficulty comes from the fact that the kernel $\widetilde{k}_t$ is generally unbounded, preventing a direct application of classical results.

The classical KRR analysis assumes a bounded kernel. Since the conditional Stein kernel is generally unbounded, we introduce a truncation argument. For a threshold $B > 0$, define the localized event

$$E_B = \{z_t \in \mathbb{R}^{2d} : \widetilde{k}_t(z_t, z_t) \le B^2\},$$

and split the covariance operator as

$$\Sigma_{\widetilde{k}_t} = \Sigma_{\le B} + \Sigma_{>B},$$

where

$$\Sigma_{\le B} = \mathbb{E}\Big[\widetilde{k}_t(\cdot, z_t) \otimes \widetilde{k}_t(\cdot, z_t)\, \mathbf{1}_{E_B}(z_t)\Big], \qquad \Sigma_{>B} = \mathbb{E}\Big[\widetilde{k}_t(\cdot, z_t) \otimes \widetilde{k}_t(\cdot, z_t)\, \mathbf{1}_{E_B^c}(z_t)\Big].$$

We also introduce the tail quantity

$$\rho_B = \mathbb{E}\Big[\widetilde{k}_t(z_t, z_t)\mathbf{1}_{E_B^c}(z_t)\Big],$$

which measures the contribution of the unbounded part of the kernel, together with

$$a_B = 1 + \frac{B^2}{n\lambda}.$$

Finally, for any $r \in L^2(\pi_{z_t})$, define

$$\tau_B(r) = \mathbb{E}\big[r(z_t)^2 \mathbf{1}_{E_B^c}(z_t)\big],$$

which satisfies $\tau_B(r) \to 0$ as $B \to \infty$ by dominated convergence.

The next two lemmas provide error bounds respectively for RKHS targets and arbitrary $L^2(\pi_{z_t})$ targets. The proofs are inspired by Bach (2024, Section 7.6).

Lemma D.11. *For every $g \in \widetilde{\mathcal{H}}_t$,*

$$\mathbb{E}\|\mathcal{S}_{n,\lambda} g - g\|^2_{L^2(\pi_{z_t})} \le \lambda a_B^2 \langle g, (\Sigma_{\widetilde{k}_t} + \lambda I)^{-1} \Sigma_{\widetilde{k}_t} g\rangle_{\widetilde{\mathcal{H}}_t} + \rho_B \|g\|^2_{\widetilde{\mathcal{H}}_t}. \tag{29}$$

In particular,

$$\mathbb{E}\|\mathcal{S}_{n,\lambda} g - g\|^2_{L^2(\pi_{z_t})} \le (\lambda a_B^2 + \rho_B)\|g\|^2_{\widetilde{\mathcal{H}}_t}. \tag{30}$$

Proof. For $g \in \widetilde{\mathcal{H}}_t$, $\frac{1}{n}\sum_{i=1}^n g(z_t^{(i)})\widetilde{k}_t(., z_t^{(i)}) = \widehat{\Sigma}_{\widetilde{k}_t} g$. Hence

$$\mathcal{S}_{n,\lambda} g = (\widehat{\Sigma}_{\widetilde{k}_t} + \lambda I)^{-1}\widehat{\Sigma}_{\widetilde{k}_t} g = (\widehat{\Sigma}_{\widetilde{k}_t} + \lambda I)^{-1}((\widehat{\Sigma}_{\widetilde{k}_t} + \lambda I) - \lambda I)g = g - \lambda(\widehat{\Sigma}_{\widetilde{k}_t} + \lambda I)^{-1} g.$$

Thus

$$\mathcal{S}_{n,\lambda} g - g = -\lambda(\widehat{\Sigma}_{\widetilde{k}_t} + \lambda I)^{-1} g.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\|\mathcal{S}_{n,\lambda} g - g\|^2_{L^2(\pi_{z_t})} = \lambda^2 \mathbb{E}\langle g, (\widehat{\Sigma}_{\widetilde{k}_t} + \lambda I)^{-1} \Sigma_{\widetilde{k}_t} (\widehat{\Sigma}_{\widetilde{k}_t} + \lambda I)^{-1} g\rangle_{\widetilde{\mathcal{H}}_t} \\
&= \lambda^2 \mathbb{E}\langle g, (\widehat{\Sigma}_{\widetilde{k}_t} + \lambda I)^{-1} \Sigma_{\le B} (\widehat{\Sigma}_{\widetilde{k}_t} + \lambda I)^{-1} g\rangle_{\widetilde{\mathcal{H}}_t} + \lambda^2 \mathbb{E}\langle g, (\widehat{\Sigma}_{\widetilde{k}_t} + \lambda I)^{-1} \Sigma_{>B} (\widehat{\Sigma}_{\widetilde{k}_t} + \lambda I)^{-1} g\rangle_{\widetilde{\mathcal{H}}_t}.
\end{aligned}$$

For the bounded part, introduce an independent sample $z_t^{(n+1)}$ and set

$$\phi_i = \widetilde{k}_t(\cdot, z_t^{(i)}), \qquad A = \widehat{\Sigma}_{\widetilde{k}_t} + \lambda I, \qquad C_{n+1} = \frac{1}{n}\sum_{i=1}^{n+1} \phi_i \otimes \phi_i.$$

The normalization remains $1/n$: adding the extra sample does not change the regularization term $n\lambda I$ in the unnormalised system. Thus $C_{n+1} + \lambda I = A + n^{-1}\phi_{n+1} \otimes \phi_{n+1}$. Write $R = (C_{n+1} + \lambda I)^{-1}$ and $c = n^{-1}\langle \phi_{n+1}, A^{-1}\phi_{n+1}\rangle_{\widetilde{\mathcal{H}}_t}$. The rank-one inverse identity gives

$$A^{-1}\phi_{n+1} = (1 + c) R \phi_{n+1}.$$

On $E_B(z_t^{(n+1)})$, we have $0 \le c \le B^2/(n\lambda)$, so $1 + c \le a_B$. Denoting the bounded contribution by Q_B, independence of the extra sample yields

$$\begin{aligned} Q_B &:= \lambda^2 \mathbb{E}\langle g, A^{-1}\Sigma_{\le B}A^{-1}g\rangle_{\widetilde{\mathcal{H}}_t} \\ &= \lambda^2 \mathbb{E}\left[\langle g, A^{-1}\phi_{n+1}\rangle^2_{\widetilde{\mathcal{H}}_t} \mathbb{1}_{E_B}(z_t^{(n+1)})\right] \\ &\le \lambda^2 a_B^2 \mathbb{E}\left[\langle g, R\phi_{n+1}\rangle^2_{\widetilde{\mathcal{H}}_t} \mathbb{1}_{E_B}(z_t^{(n+1)})\right] \\ &\le \lambda^2 a_B^2 \mathbb{E}\langle g, R\phi_{n+1}\rangle^2_{\widetilde{\mathcal{H}}_t}. \end{aligned}$$

Since R is symmetric in all $n+1$ samples, exchangeability gives

$$\begin{aligned} \mathbb{E}\langle g, R\phi_{n+1}\rangle^2_{\widetilde{\mathcal{H}}_t} &= \frac{1}{n+1}\mathbb{E}\sum_{i=1}^{n+1}\langle g, R\phi_i\rangle^2_{\widetilde{\mathcal{H}}_t} \\ &= \frac{n}{n+1}\mathbb{E}\langle g, RC_{n+1}Rg\rangle_{\widetilde{\mathcal{H}}_t}. \end{aligned}$$

By spectral calculus, $RC_{n+1}R \preceq \lambda^{-1}(C_{n+1}+\lambda I)^{-1}C_{n+1}$. Writing $\kappa = (n+1)/n$ and $\Sigma = \Sigma_{\widetilde{k}_t}$, we therefore obtain

$$Q_B \le \frac{\lambda a_B^2}{\kappa}\left\langle g, \left(I - \lambda\mathbb{E}(C_{n+1}+\lambda I)^{-1}\right) g\right\rangle_{\widetilde{\mathcal{H}}_t}.$$

Since $\mathbb{E}C_{n+1} = \kappa\Sigma$, Jensen's inequality for inversion gives

$$\mathbb{E}(C_{n+1}+\lambda I)^{-1} \succeq (\kappa\Sigma + \lambda I)^{-1}.$$

For completeness, this follows by applying the variational identity $\langle u, M^{-1}u\rangle = \sup_v\{2\langle u, v\rangle - \langle v, Mv\rangle\}$ to $M = C_{n+1}+\lambda I$ and interchanging expectation and supremum in the inequality direction, for each fixed $u \in \widetilde{\mathcal{H}}_t$. Consequently,

$$\begin{aligned} Q_B &\le \frac{\lambda a_B^2}{\kappa}\left\langle g, \left(I - \lambda(\kappa\Sigma+\lambda I)^{-1}\right) g\right\rangle_{\widetilde{\mathcal{H}}_t} \\ &= \lambda a_B^2 \langle g, (\kappa\Sigma+\lambda I)^{-1}\Sigma g\rangle_{\widetilde{\mathcal{H}}_t} \\ &\le \lambda a_B^2 \langle g, (\Sigma+\lambda I)^{-1}\Sigma g\rangle_{\widetilde{\mathcal{H}}_t}. \end{aligned}$$

The last inequality follows by spectral calculus from $s/(\kappa s+\lambda) \le s/(s+\lambda)$ for $s \ge 0$ and $\kappa \ge 1$.

For the tail part, use $\Sigma_{>B} \preceq \rho_B I$ and $(\widehat{\Sigma}_{\widetilde{k}_t}+\lambda I)^{-1} \preceq \lambda^{-1}I$:

$$\begin{aligned} \lambda^2 \mathbb{E}\langle g, (\widehat{\Sigma}_{\widetilde{k}_t}+\lambda I)^{-1}\Sigma_{>B}(\widehat{\Sigma}_{\widetilde{k}_t}+\lambda I)^{-1}g\rangle_{\widetilde{\mathcal{H}}_t} &\le \lambda^2\rho_B\mathbb{E}\left\|(\widehat{\Sigma}_{\widetilde{k}_t}+\lambda I)^{-1}g\right\|^2_{\widetilde{\mathcal{H}}_t} \\ &\le \rho_B\|g\|^2_{\widetilde{\mathcal{H}}_t}. \end{aligned}$$

This proves (29). Since $0 \preceq (\Sigma_{\widetilde{k}_t}+\lambda I)^{-1}\Sigma_{\widetilde{k}_t} \preceq I$, (30) follows. □

Lemma D.12. *For every $r \in L^2(\pi_{z_t})$,*

$$\mathbb{E}\|\mathcal{S}_{n,\lambda}r - r\|^2_{L^2(\pi_{z_t})} \le a_B^2\|r\|^2_{L^2(\pi_{z_t})} + 2\tau_B(r) + \frac{2\rho_B}{\lambda}\|r\|^2_{L^2(\pi_{z_t})}. \tag{31}$$

Consequently,

$$\mathbb{E}\|\mathcal{S}_{n,\lambda}r - r\|^2_{L^2(\pi_{z_t})} \le 3\left(a_B^2 + \frac{\rho_B}{\lambda}\right)\|r\|^2_{L^2(\pi_{z_t})}. \tag{32}$$

Proof. Let $r_i := r(z_t^{(i)})$ and define

$$G_n := \sum_{i=1}^n \widetilde{k}_t(.,z_t^{(i)}) \otimes \widetilde{k}_t(.,z_t^{(i)}) + n\lambda I, \qquad b_n := \sum_{i=1}^n r_i\widetilde{k}_t(.,z_t^{(i)}).$$

With this notation, we have $\mathcal{S}_{n,\lambda}r = G_n^{-1}b_n$. Let $z_t^{(n+1)}$ be an independent ghost point, and define also

$$\widetilde{G} := G_n + \widetilde{k}_t(.,z_t^{(n+1)}) \otimes \widetilde{k}_t(.,z_t^{(n+1)}), \qquad \widetilde{b} := b_n + r_{n+1}\widetilde{k}_t(.,z_t^{(n+1)}),$$

and $\widetilde{r} := \widetilde{G}^{-1}\widetilde{b} \in \widetilde{\mathcal{H}}_t$. Equivalently, $\widetilde{r}$ is the KRR solution on the $n+1$ samples with regularisation parameter $n\lambda$ in the unnormalised objective:

$$\widetilde{r} = \arg\min_{g \in \widetilde{\mathcal{H}}_t} \left\{ \sum_{i=1}^{n+1} (g(z_t^{(i)}) - r_i)^2 + n\lambda \|g\|^2_{\widetilde{\mathcal{H}}_t} \right\}. \tag{33}$$

Let

$$s := \mathcal{S}_{n,\lambda} r(z_t^{(n+1)}) = \langle G_n^{-1} b_n, \widetilde{k}_t(., z_t^{(n+1)}) \rangle_{\widetilde{\mathcal{H}}_t}, \quad c := \langle \widetilde{k}_t(., z_t^{(n+1)}), G_n^{-1} \widetilde{k}_t(., z_t^{(n+1)}) \rangle_{\widetilde{\mathcal{H}}_t}.$$

We can show that

$$\widetilde{r}(z_t^{(n+1)}) = s + \frac{c}{1+c}(r_{n+1} - s) = \frac{s + c r_{n+1}}{1+c}.$$

Therefore, $s - r_{n+1} = (1+c)(\widetilde{r}(z_t^{(n+1)}) - r_{n+1})$. On the localised event $E_B(z_t^{(n+1)})$, since $G_n \succeq n\lambda I$,

$$c \le \frac{\widetilde{k}_t(z_t^{(n+1)}, z_t^{(n+1)})}{n\lambda} \le \frac{B^2}{n\lambda},$$

so $1 + c \le a_B$. We split the population test error according to the ghost event. First,

$$\mathbb{E}\left[(\mathcal{S}_{n,\lambda} r(z_t^{(n+1)}) - r_{n+1})^2 \mathbb{1}_{E_B}(z_t^{(n+1)}) \right] \le a_B^2 \mathbb{E}\left[(\widetilde{r}(z_t^{(n+1)}) - r_{n+1})^2 \right].$$

By exchangeability of $z_t^{(1)}, \dots, z_t^{(n+1)}$,

$$\mathbb{E}\left[(\widetilde{r}(z_t^{(n+1)}) - r_{n+1})^2 \right] = \mathbb{E}\left[\frac{1}{n+1} \sum_{i=1}^{n+1} (\widetilde{r}(z_t^{(i)}) - r_i)^2 \right].$$

Testing the objective (33) at $g = 0$ gives

$$\sum_{i=1}^{n+1} (\widetilde{r}(z_t^{(i)}) - r_i)^2 + n\lambda \|\widetilde{r}\|^2_{\widetilde{\mathcal{H}}_t} \le \sum_{i=1}^{n+1} r_i^2.$$

Then we have $\mathbb{E}\left[(\widetilde{r}(z_t^{(n+1)}) - r_{n+1})^2 \right] \le \|r\|^2_{L^2(\pi_{z_t})}$, and thus

$$\mathbb{E}\left[(\mathcal{S}_{n,\lambda} r(z_t^{(n+1)}) - r_{n+1})^2 \mathbb{1}_{E_B}(z_t^{(n+1)}) \right] \le a_B^2 \|r\|^2_{L^2(\pi_{z_t})}. \tag{34}$$

On the tail event $E_B^c(z_t^{(n+1)})$, use $(u-v)^2 \le 2u^2 + 2v^2$:

$$\mathbb{E}\left[(\mathcal{S}_{n,\lambda} r(z_t^{(n+1)}) - r_{n+1})^2 \mathbb{1}_{E_B^c}(z_t^{(n+1)}) \right] \le 2\tau_B(r) + 2\mathbb{E}\left[\mathcal{S}_{n,\lambda} r(z_t^{(n+1)})^2 \mathbb{1}_{E_B^c}(z_t^{(n+1)}) \right].$$

Conditioning on $z_t^{(1)}, \dots, z_t^{(n)}$ and using the tail operator bound,

$$\mathbb{E}_{z_t^{(n+1)}}\left[\mathcal{S}_{n,\lambda} r(z_t^{(n+1)})^2 \mathbb{1}_{E_B^c}(z_t^{(n+1)}) \mid z_t^{(1)}, \dots, z_t^{(n)} \right] = \langle \mathcal{S}_{n,\lambda} r, \Sigma_{>B} \mathcal{S}_{n,\lambda} r \rangle_{\widetilde{\mathcal{H}}_t} \le \rho_B \|\mathcal{S}_{n,\lambda} r\|^2_{\widetilde{\mathcal{H}}_t}.$$

It remains to bound the RKHS norm of $\mathcal{S}_{n,\lambda} r$. Since $\mathcal{S}_{n,\lambda} r$ minimises

$$g \mapsto \sum_{i=1}^{n} (g(z_t^{(i)}) - r_i)^2 + n\lambda \|g\|^2_{\widetilde{\mathcal{H}}_t},$$

testing against $g = 0$ gives $n\lambda \|\mathcal{S}_{n,\lambda} r\|^2_{\widetilde{\mathcal{H}}_t} \le \sum_{i=1}^n r_i^2$. Thus $\mathbb{E}\|\mathcal{S}_{n,\lambda} r\|^2_{\widetilde{\mathcal{H}}_t} \le \frac{\|r\|^2_{L^2(\pi_{z_t})}}{\lambda}$. Consequently,

$$\mathbb{E}\left[(\mathcal{S}_{n,\lambda} r(z_t^{(n+1)}) - r_{n+1})^2 \mathbb{1}_{E_B^c}(z_t^{(n+1)}) \right] \le 2\tau_B(r) + \frac{2\rho_B}{\lambda} \|r\|^2_{L^2(\pi_{z_t})}. \tag{35}$$

Adding (34) and (35) proves Equation (31). Since $a_B \ge 1$ and $\tau_B(r) \le \|r\|^2_{L^2(\pi_{z_t})}$, (32) follows. $\square$

D.3 Proof of Theorem 5.6

Recall that $\widetilde{\mathcal{H}}_t = S_{\vec{\pi}_{t|\star}}\mathcal{H}^d = \{S_{\vec{\pi}_{t|\star}}v,\ v \in \mathcal{H}^d\}$ is an RKHS of kernel $\widetilde{k}_t$ given in Proposition D.1. In this section we use the notation $z_t = (\vec{X}_t, X_\star) \in \mathbb{R}^{2d}$ and π_{z_t} for the joint distribution. Here, Ψ_t denotes the Stein target defined above. Let $\Psi_t = h_t + h_t^\perp$ be the orthogonal decomposition of Ψ_t in $L^2(\pi_{z_t}) = \overline{\widetilde{\mathcal{H}}_t} \overset{\perp}{\oplus} \widetilde{\mathcal{H}}_t^\perp$.

The proof is composed of four steps. We first reformulate the problem as a kernel ridge regression problem. We then decompose the excess risk into an RKHS approximation term and an orthogonal component. These two terms are bounded separately before choosing the truncation level.

Step 1. By Proposition D.9,

$$R_t^c(\hat{v}_\lambda) = \|\hat{g}_\lambda - \Psi_t\|^2_{L^2(\pi_{z_t})},$$

where $\hat{g}_\lambda = \mathcal{S}_{n,\lambda}\Psi_t$, and $\mathcal{S}_{n,\lambda}$ denotes the kernel ridge regression operator introduced in Section D.2.

Step 2. Using the orthogonal decomposition $\Psi_t = h_t + h_t^\perp$, we obtain

$$\begin{aligned}
\mathbb{E}\, R_t^c(\hat{v}_{t,\lambda}) &= \mathbb{E}\, \|\hat{g}_\lambda - \Psi_t\|^2_{L^2(\pi_{z_t})} \\
&= \mathbb{E}\, \|\mathcal{S}_{n,\lambda}\Psi_t - \Psi_t\|^2_{L^2(\pi_{z_t})} \\
&= \mathbb{E}\, \left\|\mathcal{S}_{n,\lambda}h_t - h_t + \mathcal{S}_{n,\lambda}h_t^\perp - h_t^\perp\right\|^2_{L^2(\pi_{z_t})} \qquad (36)\\
&= \mathbb{E}\, \left[\left\|\mathcal{S}_{n,\lambda}h_t - h_t + \mathcal{S}_{n,\lambda}h_t^\perp\right\|^2_{L^2(\pi_{z_t})}\right] + \left\|h_t^\perp\right\|^2_{L^2(\pi_{z_t})}. \qquad (37)
\end{aligned}$$

The passage from (36) to (37) uses orthogonality to the $L^2(\pi_{z_t})$-closure of $\widetilde{\mathcal{H}}_t$. Indeed, almost surely both $\mathcal{S}_{n,\lambda}h_t$ and $\mathcal{S}_{n,\lambda}h_t^\perp$ belong to $\widetilde{\mathcal{H}}_t$, while $h_t \in \overline{\widetilde{\mathcal{H}}_t}$. Hence $\mathcal{S}_{n,\lambda}h_t - h_t + \mathcal{S}_{n,\lambda}h_t^\perp$ lies in $\overline{\widetilde{\mathcal{H}}_t}$ and is orthogonal to $h_t^\perp \in (\overline{\widetilde{\mathcal{H}}_t})^\perp$.

By Proposition D.7 we directly have that $\inf_{v \in W^{1,2}(\vec{\pi}_t)} R_t^c(v) = \|h_t^\perp\|^2_{L^2(\pi_{z_t})}$, so the term $\|h_t^\perp\|^2_{L^2(\pi_{z_t})}$ is cancelled by $\inf_{v \in W^{1,2}(\vec{\pi}_t)} R_t^c(v)$.

Step 3 We can split the remaining term in two parts

$$E\, \left[\left\|\mathcal{S}_{n,\lambda}h_t - h_t + \mathcal{S}_{n,\lambda}h_t^\perp\right\|^2_{L^2(\pi_{z_t})}\right] \le 2\mathbb{E}\left[\|\mathcal{S}_{n,\lambda}h_t - h_t\|^2_{L^2(\pi_{z_t})}\right] + 2\mathbb{E}\left[\left\|\mathcal{S}_{n,\lambda}h_t^\perp\right\|^2_{L^2(\pi_{z_t})}\right],$$

using Young's inequality: $(a+b)^2 \le 2a^2 + 2b^2$. It remains to bound separately the two terms.

Let $B > 0$ be a truncation level and recall the notation introduced in Section D.2: $E_B = \{z_t \in \mathbb{R}^{2d} :\ \widetilde{k}_t(z_t, z_t) \le B^2\}$, $\rho_B = \mathbb{E}\big[\widetilde{k}_t(z_t, z_t)\mathbf{1}_{E_B^c}(z_t)\big]$, $a_B = 1 + \frac{B^2}{n\lambda}$ and for $r \in L^2(\pi_{z_t})$, define $\tau_B(r) = \mathbb{E}\big[r(z_t)^2\mathbf{1}_{E_B^c}(z_t)\big]$.

Step 3.1. We first control $\mathbb{E}\left[\|\mathcal{S}_{n,\lambda}h_t - h_t\|^2_{L^2(\pi_{z_t})}\right]$. The idea is to use that since h_t is in the $L^2(\pi_{z_t})$-closure of $\widetilde{\mathcal{H}}_t$, for every $\epsilon > 0$ we can find $g \in \widetilde{\mathcal{H}}_t$ such that $\|g - h_t\|_{L^2(\pi_{z_t})} \le \epsilon$. For any such $g \in \widetilde{\mathcal{H}}_t$ define $r := h_t - g$. We can decompose

$$\mathcal{S}_{n,\lambda}h_t - h_t = (\mathcal{S}_{n,\lambda}g - g) + (\mathcal{S}_{n,\lambda}r - r).$$

Applying $\|u+v\|^2 \le 2\|u\|^2 + 2\|v\|^2$, followed by Lemma D.11 and Lemma D.12 to the two terms, gives

$$\begin{aligned}
\mathbb{E}\|\mathcal{S}_{n,\lambda}h_t - h_t\|^2_{L^2(\pi_{z_t})} &\le 2(\lambda a_B^2 + \rho_B)\|g\|^2_{\widetilde{\mathcal{H}}_t} + 6\left(a_B^2 + \frac{\rho_B}{\lambda}\right)\|r\|^2_{L^2(\pi_{z_t})} \\
&\le 6\left(a_B^2 + \frac{\rho_B}{\lambda}\right)\left(\|h_t - g\|^2_{L^2(\pi_{z_t})} + \lambda\|g\|^2_{\widetilde{\mathcal{H}}_t}\right).
\end{aligned}$$

As this inequality is true for every $g \in \widetilde{\mathcal{H}}_t$ we finally get

$$\mathbb{E}\|\mathcal{S}_{n,\lambda}h_t - h_t\|^2_{L^2(\pi_{z_t})} \leq 6\left(a_B^2 + \frac{\rho_B}{\lambda}\right) \inf_{g\in\widetilde{\mathcal{H}}_t} \left\{\|h_t - g\|^2_{L^2(\pi_{z_t})} + \lambda\|g\|^2_{\widetilde{\mathcal{H}}_t}\right\}.$$

Step 3.2. It remains to bound $\mathbb{E}\left[\left\|\mathcal{S}_{n,\lambda}h_t^\perp\right\|^2_{L^2(\pi_{z_t})}\right]$. The term $\mathcal{S}_{n,\lambda}h_t^\perp$ belongs to $\widetilde{\mathcal{H}}_t$ while $h_t^\perp \in \widetilde{\mathcal{H}}_t^\perp$, by orthogonality we have

$$\mathbb{E}\left\|\mathcal{S}_{n,\lambda}h_t^\perp - h_t^\perp\right\|^2_{L^2(\pi_{z_t})} = \mathbb{E}\left[\left\|\mathcal{S}_{n,\lambda}h_t^\perp\right\|^2_{L^2(\pi_{z_t})}\right] + \left\|h_t^\perp\right\|^2_{L^2(\pi_{z_t})},$$

Applying Equation (31) in Lemma D.12 to $\mathbb{E}\left\|\mathcal{S}_{n,\lambda}h_t^\perp - h_t^\perp\right\|^2_{L^2(\pi_{z_t})}$ we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathcal{S}_{n,\lambda}h_t^\perp\right\|^2_{L^2(\pi_{z_t})}\right] &= \mathbb{E}\left\|\mathcal{S}_{n,\lambda}h_t^\perp - h_t^\perp\right\|^2_{L^2(\pi_{z_t})} - \left\|h_t^\perp\right\|^2_{L^2(\pi_{z_t})} \\
&\leq a_B^2\left\|h_t^\perp\right\|^2_{L^2(\pi_{z_t})} + 2\tau_B(h_t^\perp) + \frac{2\rho_B}{\lambda}\left\|h_t^\perp\right\|^2_{L^2(\pi_{z_t})} - \left\|h_t^\perp\right\|^2_{L^2(\pi_{z_t})} \\
&= \left(a_B^2 - 1 + \frac{2\rho_B}{\lambda}\right)\left\|h_t^\perp\right\|^2_{L^2(\pi_{z_t})} + 2\tau_B(h_t^\perp),
\end{aligned}$$

and

$$\mathbb{E}\left[\left\|\mathcal{S}_{n,\lambda}h_t^\perp\right\|^2_{L^2(\pi_{z_t})}\right] \leq 3\left(\left(a_B^2 - 1\right) + \frac{\rho_B}{\lambda}\right)\left\|h_t^\perp\right\|^2_{L^2(\pi_{z_t})} + 2\tau_B(h_t^\perp).$$

By Lemma D.8, $\left\|h_t^\perp\right\|^2_{L^2(\pi_{z_t})} \leq c_t = 2d\frac{(e^{-2t}+\alpha)^2}{\sigma_t^4} + \frac{e^{-2t}}{\sigma_t^2}\mathbb{E}[\|X_\star\|^2]$ so finally

$$\mathbb{E}\left[\left\|\mathcal{S}_{n,\lambda}h_t^\perp\right\|^2_{L^2(\pi_{z_t})}\right] \leq 3\left(\left(a_B^2 - 1\right) + \frac{\rho_B}{\lambda}\right)c_t + 2\tau_B(h_t^\perp).$$

Combining all the steps we get for every $B > 0$

$$\begin{aligned}
\mathbb{E}\left[R_t^c(\hat{v}_{t,\lambda}) - \inf_{v\in W^{1,2}(\vec{\pi}_t)} R_t^c(v)\right] \leq 12\left(a_B^2 + \frac{\rho_B}{\lambda}\right) \inf_{g\in\widetilde{\mathcal{H}}_t} \left\{\|h_t - g\|^2_{L^2(\pi_{z_t})} + \lambda\|g\|^2_{\widetilde{\mathcal{H}}_t}\right\} \\
+ 6\left(\left(a_B^2 - 1\right) + \frac{\rho_B}{\lambda}\right)c_t + 4\tau_B(h_t^\perp).
\end{aligned} \tag{38}$$

Step 4. It remains to choose the truncation level B. By Cauchy–Schwarz,

$$\rho_B \leq \mathbb{E}[\widetilde{k}_t(z_t, z_t)^2]^{1/2}\mathbb{P}(E_B^c)^{1/2}.$$

By Hölder's inequality, with $q > 1$ from Assumption 5.5,

$$\tau_B(h_t^\perp) \leq \mathbb{E}\left[|h_t^\perp(z_t)|^{2q}\right]^{1/q}\mathbb{P}(E_B^c)^{(q-1)/q} = \|h_t^\perp\|^2_{L^{2q}(\pi_{z_t})}\mathbb{P}(E_B^c)^{(q-1)/q}.$$

For $0 < \lambda \leq e^{-1}$, choose

$$B_\lambda^2 = \frac{12dqM}{\sigma_t^2(q-1)}\log(1/\lambda).$$

Then Lemma D.3 yields

$$\begin{aligned}
\mathbb{P}(E_{B_\lambda}^c) &\leq \exp(d)\exp\left[-\frac{\sigma_t^2}{3}\left(\frac{B_\lambda^2}{2dM} - 1\right)\right] \\
&= \exp\left(d + \sigma_t^2/3\right)\lambda^{2q/(q-1)}.
\end{aligned}$$

Moreover, the second moment estimate in Lemma D.3 gives

$$\mathbb{E}[\widetilde{k}_t(z_t, z_t)^2]^{1/2} \leq \sqrt{8}\, dM\left(1 + \frac{d(d+2)}{\sigma_t^4}\right)^{1/2} < \infty.$$

Consequently, for fixed $t > 0$ and fixed problem parameters,

$$\frac{\rho_{B_\lambda}}{\lambda} \lesssim \lambda^{1/(q-1)}, \qquad \tau_{B_\lambda}(h_t^\perp) \lesssim \lambda^2,$$

where the implicit constants do not depend on n or λ. Finally,

$$a_{B_\lambda} = 1 + \frac{12dqM}{\sigma_t^2(q-1)} \frac{\log(1/\lambda)}{\lambda n}.$$

Writing $c_{t,q} = 12dqM/(\sigma_t^2(q-1))$, we obtain

$$\begin{aligned} a_{B_\lambda}^2 - 1 &= c_{t,q} \frac{\log(1/\lambda)}{\lambda n} \left(2 + c_{t,q} \frac{\log(1/\lambda)}{\lambda n}\right) \\ &\lesssim \frac{\log(1/\lambda)}{\lambda n} \left(1 + \frac{\log(1/\lambda)}{\lambda n}\right). \end{aligned}$$

Thus the larger constants in Lemma D.3 leave the powers of n and λ in the resulting bound unchanged. Finally, by applying the above to (38),

$$\begin{aligned} &\mathbb{E}\left[R_t^c(\hat{v}_{t,\lambda}) - \inf_{v \in W^{1,2}(\vec{\pi}_t)} R_t^c(v)\right] \\ &\lesssim \left(\left(1 + \frac{\log(1/\lambda)}{\lambda n}\right)^2 + \lambda^{1/(q-1)}\right) \inf_{g \in \widetilde{\mathcal{H}}_t} \left\{\|h_t - g\|_{L^2(\pi_{z_t})}^2 + \lambda \|g\|_{\widetilde{\mathcal{H}}_t}^2\right\} \\ &+ \frac{\log(1/\lambda)}{\lambda n} \left(2 + \frac{\log(1/\lambda)}{\lambda n}\right) + \left(\lambda^{1/(q-1)} + \lambda^2\right). \end{aligned}$$

D.4 Convergence rate: Proof of Corollary 5.7

Recall that $T : L^2(\pi_{z_t}) \to L^2(\pi_{z_t})$, $u \to \int u(z_t - t)\widetilde{k}_t(., z_t - t)\mathrm{d}\pi_{z_t}(z_t)$ is the integral operator in $L^2(\pi_{z_t})$ associated with $\widetilde{k}_t$. For $r > 0$, $[\widetilde{\mathcal{H}}_t]^r = \text{Range}(T^r)$ so if $g \in [\widetilde{\mathcal{H}}_t]^r$, there exists $u \in L^2(\pi_{z_t})$ such that $g = T^r u$.

To prove Corollary 5.7, it remains to control the term

$$A(h_t, \lambda) = \inf_{g \in \widetilde{\mathcal{H}}_t} \left\{\|h_t - g\|_{L^2(\pi_{z_t})}^2 + \lambda \|g\|_{\widetilde{\mathcal{H}}_t}^2\right\}.$$

To this end, we assume that h_t belongs to an interpolation space between $\widetilde{\mathcal{H}}_t$ and $L^2(\pi_{z_t})$.

Lemma D.13. *Suppose that for $0 < r \leq \frac{1}{2}$, $h_t \in [\widetilde{\mathcal{H}}_t]^r$. Then*

$$A(h_t, \lambda) = O(\lambda^{2r}).$$

Proof. Let $(e_j)_{j \geq 1}$ be an orthonormal sequence in $L^2(\pi_{z_t})$ and $(\mu_j)_{j \geq 1}$ a sequence in $\mathbb{R}^+$ such that $T = \sum_{j \geq 1} \mu_j e_j e_j^*$. Writing $h_t = \sum_{j \geq 1} h_j e_j$ and $g = \sum_{j \geq 1} g_j e_j$, we have $\|g\|_{\widetilde{\mathcal{H}}_t}^2 = \sum_{j \geq 1} \frac{g_j^2}{\mu_j}$ and $\|h_t - g\|_{L^2(\pi_{z_t})}^2 = \sum_{j \geq 1} (h_j - g_j)^2$. Therefore,

$$A(h_t, \lambda) = \inf_{(g_j)} \sum_{j \geq 1} \left((h_j - g_j)^2 + \lambda \frac{g_j^2}{\mu_j}\right).$$

The minimiser can be computed for each j and is $g_j = \frac{\mu_j}{\mu_j + \lambda} h_j$. By replacing g_j in the previous equation we get

$$A(h_t, \lambda) = \sum_{j \geq 1} \frac{\lambda}{\mu_j + \lambda} h_j^2.$$

Since $h_t = T^r u$, we have $h_j = \mu_j^r u_j$ where we write $u = \sum_{j \geq 1} u_i e_i$. Hence

$$A(h_t, \lambda) = \sum_{j \geq 1} \frac{\lambda \mu_j^{2r}}{\mu_j + \lambda} u_j^2.$$

With $0 < r \leq 1/2$, $\frac{\lambda \mu_j^{2r}}{\mu_j + \lambda} = \lambda^{2r} \frac{(\mu_j/\lambda)^{2r}}{1+\mu_j/\lambda} \leq C_r \lambda^{2r}$, where $C_r = \sup_{t \geq 0} \frac{t^{2r}}{1+t} < \infty$. Therefore,

$$A(h_t, \lambda) \leq C_r \lambda^{2r} \|u\|^2_{L^2(\pi_{z_t})}.$$

This proves the result. □

Recalling that the bound of Theorem 5.6 is

$$\begin{aligned}
&\mathbb{E}\left[R_t^c(\hat{v}_{t,\lambda}) - \inf_{v \in W^{1,2}(\vec{\pi}_t)} R_t^c(v)\right] \\
&\lesssim \underbrace{\left(\left(1 + \frac{\log(1/\lambda)}{\lambda n}\right)^2 + \lambda^{1/(q-1)}\right) \inf_{g \in \widetilde{\mathcal{H}}_t} \left\{\|h_t - g\|^2_{L^2(\pi_{z_t})} + \lambda \|g\|^2_{\widetilde{\mathcal{H}}_t}\right\}}_{=A} \\
&+ \underbrace{\frac{\log(1/\lambda)}{\lambda n}\left(2 + \frac{\log(1/\lambda)}{\lambda n}\right)}_{=B} + \underbrace{\left(\lambda^{1/(q-1)} + \lambda^2\right)}_{=C},
\end{aligned}$$

we have $A = O(\lambda^{2r})$, $B = O\left(\frac{\log(1/\lambda)}{\lambda n}\right)$ and $C = O\left(\max\{\lambda^2, \lambda^{1/(q-1)}\}\right)$ when $\lambda \to 0$. We can easily suppose that $q \leq 2$. Indeed if Assumption 5.5 is satisfied for $q > 2$ then it is also satisfied for any $q' < q$ by the inclusion of $L^{2q}(\pi_{z_t})$ in $L^{2q'}(\pi_{z_t})$. Under this hypothesis, $\lambda^{1/(q-1)} = O(\lambda)$ and so $C = O(\lambda)$.

If we choose $\lambda = \frac{1}{n^{1/2}}$, the slowest rate is among $\frac{1}{n^r}$, $\frac{\log(n)}{n^{1/2}}$ and $\frac{1}{n^{\frac{1}{2}}}$. As $r < \frac{1}{2}$, the final rate is

$$\mathbb{E}\left[R_t^c(\hat{v}_{t,\lambda}) - \inf_{v \in W^{1,2}(\vec{\pi}_t)} R_t^c(v)\right] = O\left(\frac{1}{n^r}\right).$$

D.5 Comparing rates of convergence with kernel score matching

In this section, we establish a result similar to Corollary 5.7 for the convergence rate of the vector field estimator which shows that the score-matching kernel estimator and the Stein-Transport kernel estimator have equivalent statistical performance. Let $\hat{s}_{t,\lambda} = (s_t^{1,*}, \ldots, s_t^{d,*})$ collect the coordinate minimisers of (22).

Theorem D.14. *Suppose there exists $R > 0$ such that $\|k(., x)\|_{\mathcal{H}} \leq R$ for all $x \in \mathbb{R}^d$. Then,*

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{L}_c^{SM}(\hat{s}_{t,\lambda}) - \inf_{s \in L^2(\vec{\pi}_t)} \mathcal{L}_c^{SM}(s)\right] \\
&\quad \leq \frac{d}{\sigma_t^2} \frac{R^2}{\lambda n}\left(2 + \frac{R^2}{\lambda n}\right) \\
&\qquad + 3\left(1 + \frac{R^2}{\lambda n}\right)^2 \inf_{s \in \mathcal{H}^d} \left\{\|s - s^*\|^2_{L^2(\vec{\pi}_t)} + \lambda \|s\|^2_{\mathcal{H}^d}\right\}.
\end{aligned} \tag{39}$$

Proof. The proof for Score Matching follows a similar outline to that of Stein Transport. We will therefore skip over certain details and focus only on the points that differ between the two proofs. For clarity, the step numbers correspond to those in Stein Transport. Since the kernel is bounded in Score Matching, there is no Step 4 in this proof. Unlike in Theorem 5.6, where we integrated over the pair $z_t = (\vec{X}_t, X_\star)$, here we will use the pair $(\vec{X}_t, \varepsilon_t)$, which is more convenient here, and we write $\pi_{\vec{X}_t, \varepsilon_t}$ for the coupling distribution and $\vec{\pi}_{t|\varepsilon_t}$ for the conditional distribution of $\vec{X}_t | \varepsilon_t$.

Step 1. Write $Y_t = -\varepsilon_t/\sigma_t = \nabla \log \vec{\pi}_{t|\star}(\vec{X}_t | X_\star)$. The denoising identity gives $s^*(\vec{X}_t) = \mathbb{E}[Y_t \mid \vec{X}_t] = \nabla \log \vec{\pi}_t(\vec{X}_t)$ Vincent (2011). Thus $Y_t - s^*(\vec{X}_t)$ is orthogonal to every square-integrable

function of $\vec{X}_t$, and

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}_c^{SM}(\hat{s}_{t,\lambda}) - \inf_{s\in L^2(\vec{\pi}_t)} \mathcal{L}_c^{SM}(s)\right] \\ = \mathbb{E}\left[\|\hat{s}_{t,\lambda} - Y_t\|^2_{L^2(\pi_{\vec{X}_t,\varepsilon_t})} - \|s^* - Y_t\|^2_{L^2(\pi_{\vec{X}_t,\varepsilon_t})}\right] \\ = \mathbb{E}\|\hat{s}_{t,\lambda} - s^*\|^2_{L^2(\vec{\pi}_t)}.\end{aligned}$$

Step 2. Let $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n k(\cdot, \vec{X}_t^{(i)}) \otimes k(\cdot, \vec{X}_t^{(i)})$. Define $\mathcal{T}_{n,\lambda} : L^2(\pi_{\vec{X}_t,\varepsilon_t}) \to L^2(\vec{\pi}_t)$ such that for $s \in L^2(\pi_{\vec{X}_t,\varepsilon_t})$, $\mathcal{T}_{n,\lambda}s = (\hat{\Sigma} + \lambda I)^{-1}\frac{1}{n}\sum_{i=1}^n s(\vec{X}_t^{(i)}, \varepsilon_t^{(i)})k(., \vec{X}_t^{(i)})$. The minimiser of Equation (22) is written

$$\begin{aligned}\hat{s}_{t,\lambda} &= (\hat{\Sigma} + \lambda I)^{-1}\frac{1}{n}\sum_{i=1}^n \left(-\frac{\varepsilon_t^{(i)}}{\sigma_t}\right) k(., \vec{X}_t^{(i)}) \\ &= (\hat{\Sigma} + \lambda I)^{-1}\frac{1}{n}\sum_{i=1}^n s^*(\vec{X}_t^{(i)})k(., \vec{X}_t^{(i)}) + (\hat{\Sigma} + \lambda I)^{-1}\frac{1}{n}\sum_{i=1}^n \delta(\vec{X}_t^{(i)}, \varepsilon_t^{(i)})k(., \vec{X}_t^{(i)}) \\ &= \mathcal{T}_{n,\lambda}s^* + \mathcal{T}_{n,\lambda}\delta,\end{aligned}$$

where $\delta(\vec{X}_t, \varepsilon_t) = -\varepsilon_t/\sigma_t - s^*(\vec{X}_t)$, $\mathbb{E}[\delta(\vec{X}_t, \varepsilon_t)|\vec{X}_t] = 0$ and s^* is seen as a function in $L^2(\pi_{\vec{X}_t,\varepsilon_t})$ that is constant in the variable ε_t. Then

$$\mathbb{E}\left[\|\hat{s}_{t,\lambda} - s^*\|^2_{L^2(\vec{\pi}_t)}\right] = \mathbb{E}\left[\|\mathcal{T}_{n,\lambda}s^* - s^* + \mathcal{T}_{n,\lambda}\delta\|^2_{L^2(\vec{\pi}_t)}\right].$$

Step 3. Conditionally on the training inputs $\mathcal{X} = (\vec{X}_t^{(i)})_{i=1}^n$, the term $\mathcal{T}_{n,\lambda}s^* - s^*$ is deterministic and

$$\mathbb{E}[\mathcal{T}_{n,\lambda}\delta \mid \mathcal{X}] = (\hat{\Sigma} + \lambda I)^{-1}\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\delta(\vec{X}_t^{(i)}, \varepsilon_t^{(i)}) \mid \mathcal{X}]k(., \vec{X}_t^{(i)}) = 0.$$

Consequently, the expected cross-term vanishes and

$$\mathbb{E}\|\hat{s}_{t,\lambda} - s^*\|^2_{L^2(\vec{\pi}_t)} = \mathbb{E}\|\mathcal{T}_{n,\lambda}s^* - s^*\|^2_{L^2(\vec{\pi}_t)} + \mathbb{E}\|\mathcal{T}_{n,\lambda}\delta\|^2_{L^2(\vec{\pi}_t)}.$$

Step 3.1. Bound on $\mathbb{E}\left[\|\mathcal{T}_{n,\lambda}s^* - s^*\|^2_{L^2(\vec{\pi}_t)}\right]$. Set $a = 1 + R^2/(n\lambda)$. Since $k(x,x) \le R^2$, taking $B = R$ removes both tail terms. For any $s \in \mathcal{H}^d$, apply Equation (31) to $s^* - s$ and Lemma D.11 to s, coordinate by coordinate. Using $\|u+v\|^2 \le 2\|u\|^2 + 2\|v\|^2$ gives

$$\mathbb{E}\|\mathcal{T}_{n,\lambda}s^* - s^*\|^2_{L^2(\vec{\pi}_t)} \le 2a^2\big(\|s - s^*\|^2_{L^2(\vec{\pi}_t)} + \lambda\|s\|^2_{\mathcal{H}^d}\big).$$

Taking the infimum and enlarging the numerical constant to 3 yields

$$\mathbb{E}\left[\|\mathcal{T}_{n,\lambda}s^* - s^*\|^2_{L^2(\vec{\pi}_t)}\right] \le 3\left(1 + \frac{R^2}{\lambda n}\right)^2 \inf_{s\in\mathcal{H}^d}\left\{\|s - s^*\|^2_{L^2(\vec{\pi}_t)} + \lambda\|s\|^2_{\mathcal{H}^d}\right\}.$$

Remark D.15. Here we apply Lemma D.12 and Lemma D.11 to k, $\mathcal{H}^d$ and $\mathcal{T}_{n,\lambda}$ instead of $\widetilde{k}_t$, $\widetilde{\mathcal{H}}_t$ and $\mathcal{S}_{n,\lambda}$ but we can see that the proof fits this change well.

Step 3.2 Bound on $\mathbb{E}\left[\|\mathcal{T}_{n,\lambda}\delta\|^2_{L^2(\vec{\pi}_t)}\right]$. We can equivalently bound $\mathbb{E}\left[\|\mathcal{T}_{n,\lambda}\delta\|^2_{L^2(\pi_{\vec{X}_t,\varepsilon_t})}\right]$ as $\|\mathcal{T}_{n,\lambda}\delta\|^2_{L^2(\vec{\pi}_t)} = \|\mathcal{T}_{n,\lambda}\delta\|^2_{L^2(\pi_{\vec{X}_t,\varepsilon_t})}$. We can then prove that $\mathcal{T}_{n,\lambda}\delta$ and δ are orthogonal in $L^2(\pi_{\vec{X}_t,\varepsilon_t})$:

$$\langle\mathcal{T}_{n,\lambda}\delta, \delta\rangle_{L^2(\pi_{\vec{X}_t,\varepsilon_t})} = \mathbb{E}_{\vec{X}_t}\left\langle(\mathcal{T}_{n,\lambda}\delta)(\vec{X}_t), \mathbb{E}\left[\delta(\vec{X}_t, \varepsilon_t)|\vec{X}_t\right]\right\rangle = 0.$$

So as in the proof of Theorem 5.6, $\|\mathcal{T}_{n,\lambda}\delta\|^2_{L^2(\pi_{\vec{X}_t,\varepsilon_t})} = \|\mathcal{T}_{n,\lambda}\delta - \delta\|^2_{L^2(\pi_{\vec{X}_t,\varepsilon_t})} - \|\delta\|^2_{L^2(\pi_{\vec{X}_t,\varepsilon_t})}$. Applying Equation (31) of Lemma D.12,

$$\begin{aligned}\mathbb{E}\left[\|\mathcal{T}_{n,\lambda}\delta\|^2_{L^2(\vec{\pi}_t)}\right] &= \mathbb{E}\left[\|\mathcal{T}_{n,\lambda}\delta - \delta\|^2_{L^2(\pi_{\vec{X}_t,\varepsilon_t})} - \|\delta\|^2_{L^2(\pi_{\vec{X}_t,\varepsilon_t})}\right] \\ &\leq \left(\left(1+\frac{R^2}{\lambda n}\right)^2 - 1\right)\|\delta\|^2_{L^2(\pi_{\vec{X}_t,\varepsilon_t})} = \frac{R^2}{\lambda n}\left(2+\frac{R^2}{\lambda n}\right)\|\delta\|^2_{L^2(\pi_{\vec{X}_t,\varepsilon_t})},\end{aligned}$$

and $\|\delta\|^2_{L^2(\pi_{\vec{X}_t,\varepsilon_t})} = \operatorname{Tr}\mathbb{E}_{\vec{X}_t}\left[\operatorname{Var}[\frac{\varepsilon_t}{\sigma_t}|\vec{X}_t]\right] \leq \operatorname{Tr}\operatorname{Var}\left[\frac{\varepsilon_t}{\sigma_t}\right] = \frac{d}{\sigma_t^2}$. This concludes the proof. □

Corollary D.16. *Suppose the assumptions of Theorem D.14 hold. In addition suppose that* $\nabla \log \vec{\pi}_t \in [\mathcal{H}^d]^r$ *for* $0 < r < \frac{1}{2}$. *Then, if we choose* $\lambda = \frac{1}{n^{1/2}}$,

$$\mathbb{E}\left[\mathcal{L}_c^{SM}(\hat{s}_{t,\lambda}) - \inf_{s\in L^2(\vec{\pi}_t)}\mathcal{L}_c^{SM}(s)\right] = O\left(\frac{1}{n^r}\right).$$

E Experimental details

E.1 Licenses

The following assets were used in this work.

- `NumPyro`
  Apache-2.0 License
  `https://github.com/pyro-ppl/numpyro.git`
- `ucimlrepo`
  MIT license
  `https://github.com/uci-ml-repo/ucimlrepo.git`

E.2 Experimental compute resources

The experiments were conducted on the King's College London HPC system using A100/H100 GPUs with 80 GB of allocated memory.

E.3 Computational cost

We discuss the computational cost of the proposed Stein transport methods, distinguishing the cost of fitting the velocity fields from the cost of generating new samples. Throughout, n denotes the number of training samples, d the data dimension, m the number of time-discretisation steps, and p the number of generated samples, following the notation of Algorithms 1 and 2.

Fitting cost. For standard Stein transport (DST and SST), this amounts to computing the $n \times n$ matrix Ξ_t and solving the linear system in (18). Assuming direct dense linear algebra, assembling Ξ_t requires $\mathcal{O}(n^2 d)$ operations for kernels whose values and derivatives can be evaluated in $\mathcal{O}(d)$ time, while solving the system costs $\mathcal{O}(n^3)$. Over m time steps, the total fitting cost is therefore $\mathcal{O}\big(m(n^2d+n^3)\big)$. For the vectorised variants (VDST and VSST), the regression problem decomposes into d independent scalar problems, as described in Proposition B.1. Each coordinate requires solving an $n \times n$ linear system. Using separate dense factorisations, the total fitting cost is $\mathcal{O}\big(m(n^2d+dn^3)\big)$, where the matrix assembly cost assumes that pairwise kernel quantities are shared across coordinates.

Inference latency. Sample generation proceeds by discretising the learned reverse dynamics over the same m time steps. By the representer representation in (17), evaluating the velocity field at a new point requires computing kernel values and derivatives against the n noised training samples. For kernels with $\mathcal{O}(d)$ evaluation cost, this requires $\mathcal{O}(nd)$ operations per generated sample and time step, for both standard and vectorised Stein transport. Consequently, generating p samples has an overall cost of $\mathcal{O}(mpnd)$. As a smoke test of inference latency, we conducted a small-scale timing experiment on an Apple M5 MacBook Pro. We measured the time required to generate 32 samples

for Bayesian logistic regression on the Breast Cancer dataset ($d = 30$) and the Gaussian mixture experiment ($d = 10$), both using $n = 50$ training samples and $m = 200$ time-discretisation steps. The corresponding runtimes were approximately 25 ms and 21 ms, respectively.

E.4 Details of experiments for Gaussian Mixtures and additional results

Gaussian mixture target. The target distribution is $\pi_\star := \frac{1}{4}\sum_{i=1}^{4}\mathcal{N}\big(\mu_i,\, 0.5^2 I_{10}\big)$ where the component means $\{\mu_i\}_{i=1}^{4}$ are drawn independently and uniformly from $[-5, 5]^{10}$.

OU path setup. In this experiment, we set $t_{799} = 8.0$, and use a uniformly spaced time grid $\{t_0, \ldots, t_{799}\}$.

Choice of kernel function. For SM, PF, DST, VDST, SST, and VSST, we employ a multi-scale inverse multiquadratic (IMQ) kernel

$$k_{\text{M-IMQ}}(x, y) = \frac{1}{5}\sum_{\ell=1}^{5}\left(1 + \frac{\|x - y\|^2}{\sigma_\ell^2}\right)^{-1/2}. \tag{40}$$

For the KDE baseline, we use a multi-scale Gaussian kernel and sample directly from the resulting Gaussian mixture,

$$k_{\text{M-Gaussian}}(x, y) = \frac{1}{5}\sum_{\ell=1}^{5}\exp\left(-\frac{\|x - y\|^2}{2\sigma_\ell^2}\right). \tag{41}$$

The bandwidths are selected adaptively for each $t \in [0, T]$ as

$$\sigma_\ell = s\, q_{\alpha_\ell}, \quad \alpha_\ell \in \{0.1, 0.25, 0.5, 0.75, 0.9\}, \quad \ell = 1, \ldots, 5,$$

where q_{α_ℓ} denotes the α_ℓ-quantile of the pairwise distances $\|X_t^i - X_t^j\|$ (excluding self-distances), and s is a user-specified scaling factor.

Computation of maximum mean discrepancy (MMD). We evaluate the discrepancy between synthetic samples and reference samples using the maximum mean discrepancy (MMD). Given samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, and a positive definite kernel k, the squared MMD is defined as

$$\text{MMD}^2(X, Y) = \frac{1}{n^2}\sum_{a=1}^{n}\sum_{b=1}^{n}k(X_a, X_b) - \frac{2}{nm}\sum_{a=1}^{n}\sum_{b=1}^{m}k(X_a, Y_b) + \frac{1}{m^2}\sum_{a=1}^{m}\sum_{b=1}^{m}k(Y_a, Y_b). \tag{42}$$

We report $\text{MMD} = \sqrt{\text{MMD}^2}$.

For all methods, MMD is computed using the multi-scale IMQ kernel in (40). The scaling factor is fixed to $s = 1$, and the bandwidth quantiles are determined from the pairwise distances between reference samples.

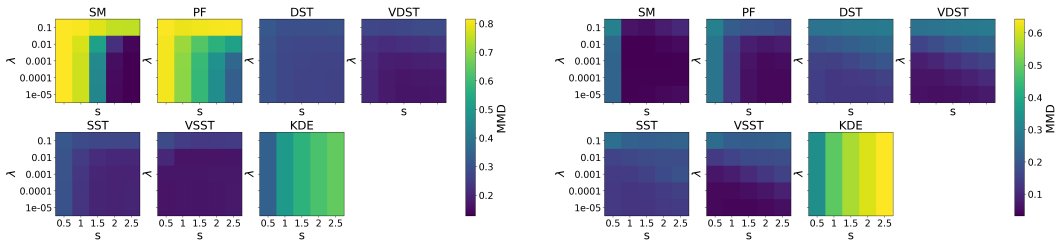
Grid search of hyperparameters. For each kernel-based method, we perform a grid search over the scaling factor and regularisation parameter,

$$s \in \{0.5, 1.0, 1.5, 2.0, 2.5\}, \quad \lambda \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\},$$

for each N_{train}.

In Figure 1, we compute the MMD between synthetic and reference samples for all hyperparameter configurations and report: (i) the minimum MMD over the grid for each training size; and (ii) the mean MMD over the grid with standard deviation error bars. For $N_{\text{train}} = 970$, we additionally show kernel density estimates of the first five marginals using the configuration achieving the lowest MMD.

To illustrate hyperparameter sensitivity, Figure 4 shows MMD heatmaps over the parameter grid for $N_{\text{train}} = 40$ and $N_{\text{train}} = 970$.



(a) $N_{\text{train}} = 40$ (b) $N_{\text{train}} = 970$

Figure 4: 10-dimensional Gaussian Mixture. MMD values over the hyperparameter grid.

E.5 Details of experiments for posterior emulation and additional results

Bayesian logistic regression setup. The dataset consists of input–label pairs $\mathcal{D} := \{(u_i, w_i)\}_{i=1}^n$, where $u_i \in \mathbb{R}^d$ denotes the covariates and $w_i \in \{0, 1\}$ the binary labels. The likelihood is

$$p(\mathcal{D} \mid x) = \prod_{i=1}^{n} \sigma(u_i^\intercal x)^{\, w_i} \big(1 - \sigma(u_i^\intercal x)\big)^{\, 1-w_i},$$

where $\sigma(z) = (1 + e^{-z})^{-1}$ is the logistic sigmoid function.

Prior to inference, the covariates are standardised componentwise to have zero mean and unit variance,

$$u_i \leftarrow \frac{u_i - \text{mean}(u)}{\text{std}(u)}, \qquad i = 1, \ldots, n.$$

Posterior samples are generated using the No-U-Turn sampler (NUTS) implemented in `NumPyro`. The first 1,000 iterations are discarded as burn-in, and the subsequent 2,000 samples are retained as the reference dataset.

OU path setup. We set $t_{499} = 8.0$ and use a uniformly spaced time grid $\{t_0, \ldots, t_{499}\}$ over $[0, 8.0]$.

Prediction accuracy. Let $\mathbf{1}(A)$ denote the indicator function defined by

$$\mathbf{1}(A) = \begin{cases} 1, & \text{if } A \text{ is true,} \\ 0, & \text{otherwise.} \end{cases}$$

Given posterior samples $X_1, \ldots, X_M$, the predicted label is computed as

$$\hat{w}_i = \mathbf{1}\left(\frac{1}{M} \sum_{a=1}^{M} \sigma(u_i^\intercal X_a) > 0.5 \right).$$

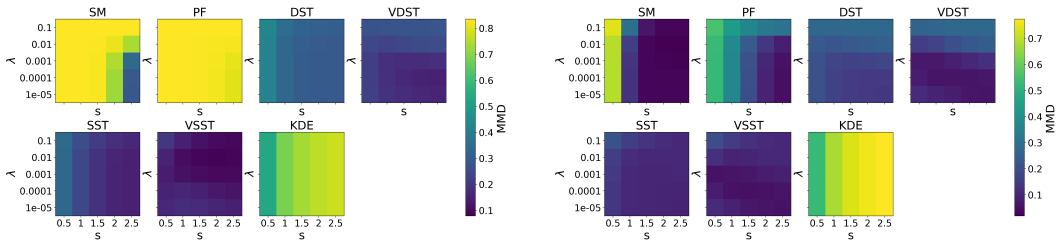
The prediction accuracy is then defined as

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(w_i = \hat{w}_i).$$

Generation of test samples. The choice of kernel, MMD computation, and hyperparameter grid search follow the setup of the 10-dimensional Gaussian mixture experiment (see Section E.4).

In Figure 2, we evaluate each method over the full hyperparameter grid and report: (i) the highest prediction accuracy and lowest MMD achieved for each training size; and (ii) the mean prediction accuracy and mean MMD over the grid, together with standard deviation error bars.

Figure 5 shows MMD heatmaps over the parameter grid for $N_{\text{train}} = 40$ and $N_{\text{train}} = 970$.



(a) $N_{\text{train}} = 40$ (b) $N_{\text{train}} = 970$

Figure 5: BLR on breast cancer data. MMD values over the hyperparameter grid.

E.6 Details of experiments for Generative Data Augmentation and additional results

Dataset details. The details of the four datasets are given in Table 2. Since the objective of this section is to perform data augmentation starting with small datasets, we used smaller samples from the California Housing (CAL) and Song Popularity (SON) datasets. The sample sizes used are listed in Table 2. For the Song Popularity dataset we removed a categorical feature, audio mode, as Stein transport only works on continuous features.

Dataset	Size	Features	Task	Source
YAC	308	6	Regression	UCI
DIA	442	10	Regression	Scikit-learn
CAL	1500	8	Regression	Scikit-learn
SON	1000	11	Regression	Kaggle

Table 2: Description of datasets

Choice and tuning of sampling methods and machine learning models. As in Kotelnikov et al. (2023), we selected CTGAN and TVAE (Xu et al., 2019) as baselines because they are well-suited for small UCI tabular datasets with purely continuous features, offering greater training stability and lower overfitting risk than other architectures, and because their code is open-source. All sampling methods have been tuned for each dataset by doing a grid search on the parameters. For VDST and SM the bandwidth scaling factor is chosen among $s \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$. The regularisation for VDST is fixed to $\lambda = 10^{-5}$ and for SM, $\lambda = 10^{-3}$. For CTGAN, embedding_dim $\in \{64, 128\}$, generator_dim $\in \{(256, 256), (128, 128)\}$, discriminator_dim $\in \{(256, 256), (128, 128)\}$, generator_lr $\in \{10^{-4}, 2 \times 10^{-4}\}$, discriminator_lr $\in \{10^{-4}, 2 \times 10^{-4}\}$, batch_size $= 100$, epochs $= 250$, and pac $\in \{5, 10\}$. For TVAE, the hyperparameters are chosen among: embedding_dim $\in \{32\}$, compress_dims $\in \{(64, 32), (128, 64)\}$, decompress_dims $\in \{(32, 64), (64, 128)\}$, batch_size $= 32$, epochs $\in \{120, 180\}$, l2scale $\in \{5 \times 10^{-6}, 10^{-5}\}$, and loss_factor $\in \{1.5, 2.5\}$.

The model to be trained to calculate the RMSE in the grid search is Random Forest. For all the models we use the scikit-learn default hyperparameters as done in Manousakas and Aydöre (2023).

Additional details and results on generative data augmentation (first experiment). Following Manousakas and Aydöre (2023), the models we use are imported from the `sklearn` library with their default hyperparameters. For VDST and SM, the iterations are done for $t_1 = 0.01, ..., t_{800} = 8.0$ with a uniform stepsize grid. Here we add the tables from the three other models, which confirm VDST performance in producing a synthetic dataset. As another measure of sampling method performance, we look at the Wasserstein distance between each synthetic training dataset and the true training dataset. We observe that in three of the four datasets, VDST has the minimum distance, while in the other dataset, SM minimizes the distance.

Table 3: RMSE Gradient Boosting

	YAC	DIA	CAL	SON
Real	0.81 ± 0.16	60.27 ± 3.58	0.42 ± 0.02	15.68 ± 0.92
VDST	**3.89 ± 1.09**	**59.07 ± 5.54**	**0.44 ± 0.01**	16.68 ± 0.91
SM	4.75 ± 1.55	60.11 ± 3.93	0.48 ± 0.03	**16.26 ± 0.83**
CTGAN	17.90 ± 2.65	82.57 ± 8.57	0.83 ± 0.07	16.81 ± 0.81
TVAE	10.27 ± 2.27	66.53 ± 6.89	0.50 ± 0.04	17.15 ± 0.64

Table 4: RMSE Linear Regression

	YAC	DIA	CAL	SON
Real	8.97 ± 1.18	55.83 ± 2.16	0.52 ± 0.03	15.50 ± 0.66
VDST	**9.14 ± 0.97**	**57.34 ± 3.23**	0.58 ± 0.04	**16.17 ± 1.18**
SM	27.70 ± 23.53	58.60 ± 2.76	**0.54 ± 0.01**	16.25 ± 0.34
CTGAN	17.30 ± 2.32	77.84 ± 8.14	0.82 ± 0.07	16.50 ± 1.00
TVAE	10.26 ± 2.33	59.83 ± 3.30	0.57 ± 0.04	18.01 ± 1.87

Table 5: RMSE Ridge Regression

	YAC	DIA	CAL	SON
Real	8.97 ± 1.19	55.88 ± 2.15	0.52 ± 0.03	15.50 ± 0.66
VDST	**9.03 ± 1.01**	**57.32 ± 3.02**	0.58 ± 0.04	**16.16 ± 1.18**
SM	27.70 ± 23.53	58.59 ± 2.76	**0.54 ± 0.01**	16.25 ± 0.34
CTGAN	17.29 ± 2.32	77.83 ± 8.13	0.82 ± 0.07	16.50 ± 1.00
TVAE	10.26 ± 2.33	59.78 ± 3.27	0.57 ± 0.04	17.09 ± 1.05

Table 6: Wasserstein distances

	YAC	DIA	CAL	SON
VDST	**11.66 ± 0.14**	**12.41 ± 0.20**	46.75 ± 0.00	**3.14 ± 0.02**
SM	43.86 ± 18.61	17.62 ± 0.36	**46.74 ± 0.01**	4.48 ± 0.13
CTGAN	13.79 ± 1.02	20.95 ± 0.43	46.79 ± 0.08	4.32 ± 0.15
TVAE	12.57 ± 0.32	15.25 ± 0.26	46.77 ± 0.01	4.06 ± 0.09

Additional details and results on generative data augmentation (second experiment). In this second experiment, for each dataset we took a subset of $N = 200$ samples. In contrast to the previous experiment, the train/test split is 50% so the real training dataset D_{train} is of size $N_{train} = 100$. Then we generated from the real training dataset, synthetic datasets of size $N_{synth} = 800$. Here iterations for VDST and SM are done from $t_1 = 0.01$ to $t_{800} = 15.0$ with constant step size between iterations. We also observe here that VDST is one of the best performing models but faces stiff competition from TVAE. However, it does not always outperform models trained on the real training dataset alone. For example, for the Yacht Hydrodynamics dataset, none of the synthetic datasets outperform any of the models on D_{train}.

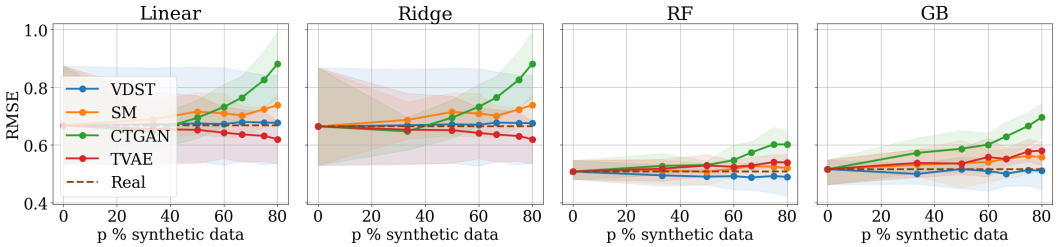


Figure 6: RMSE on (CAL) dataset.

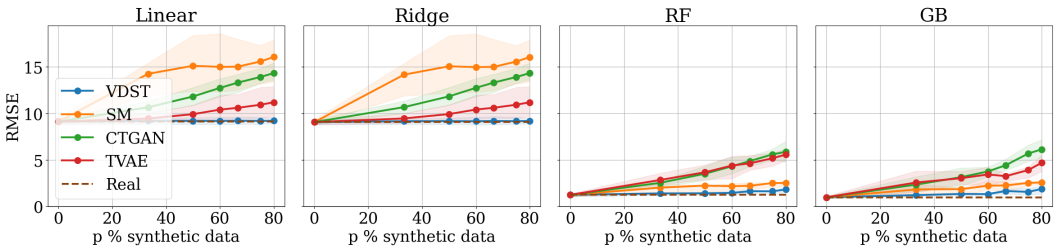


Figure 7: RMSE on (YAC) dataset.

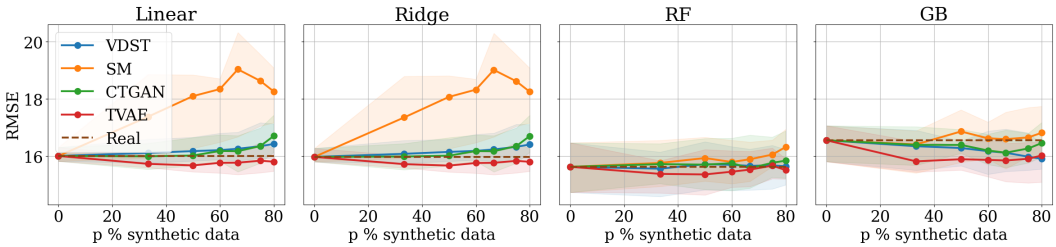


Figure 8: RMSE on (SON) dataset.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our method is developed in Section 4, with theoretical and experimental validation and support in Sections 5 and 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have included a paragraph on limitations at the end of the main part of the paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide careful, complete and rigorous proofs in Appendices A, B and D.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental details can be found in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Detailed pseudo-code is available in Appendix C, along with experimental details in Appendix E. The implementation will be released shortly.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental details are available in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We include systematic parameter sweeps that cover a wide range of hyperparameters. Across seeds, visual inspections show very little variation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include these details in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the Code of Ethics and confirm that our paper adheres to these guidelines. We see no potential harmful consequences of our work, and include extensive reproducibility details in Appendix E.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our work is mostly theory and methodology-focused. We do not see any immediate societal impacts, though certain applications may share similar societal consequences as those of generative modeling in general.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the previous works on which this research builds, and we acknowledge assets and licenses at the beginning of Appendix E.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLM assistance (proof checking and calculations for Theorem 5.6, code for the experiments of Section 6) was used after initial submission and is declared in the "Statement of AI Usage" paragraph at the end of the main text. The authors take full responsibility for all results.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.