

---

# STEIN TRANSPORT FOR BAYESIAN INFERENCE

---

A PREPRINT

Nikolas Nüsken

King's College London

[nikolas.nusken@kcl.ac.uk](mailto:nikolas.nusken@kcl.ac.uk)

September 4, 2024

## ABSTRACT

We introduce *Stein transport*, a novel methodology for Bayesian inference designed to efficiently push an ensemble of particles along a predefined curve of tempered probability distributions. The driving vector field is chosen from a reproducing kernel Hilbert space and can be derived either through a suitable kernel ridge regression formulation or as an infinitesimal optimal transport map in the Stein geometry. The update equations of Stein transport resemble those of Stein variational gradient descent (SVGD), but introduce a time-varying score function as well as specific weights attached to the particles. While SVGD relies on convergence in the long-time limit, Stein transport reaches its posterior approximation at finite time  $t = 1$ . Studying the mean-field limit, we discuss the errors incurred by regularisation and finite-particle effects, and we connect Stein transport to birth-death dynamics and Fisher-Rao gradient flows. In a series of experiments, we show that in comparison to SVGD, Stein transport not only often reaches more accurate posterior approximations with a significantly reduced computational budget, but that it also effectively mitigates the variance collapse phenomenon commonly observed in SVGD.

**Keywords** Stein variational gradient descent · Kernel ridge regression · Optimal transport

## 1 Introduction

Approximating high-dimensional probability distributions poses a key computational challenge in Bayesian inference, with Markov Chain Monte Carlo (MCMC) [Brooks et al., 2011, Robert and Casella, 2013] and Variational Inference (VI) [Jordan et al., 1999, Blei et al., 2017] representing the two most common paradigms used in practice. Combining features of both approaches, particle-based algorithms have attracted considerable interest in recent years, allowing for highly flexible as well as tractable and theoretically grounded methodologies [Liu et al., 2019, Trillos et al., 2023, Chen et al., 2023b].

A common thread is the idea of devising dynamical movement schemes that aim at reducing a suitable discrepancy towards the target, such as the Kullback-Leibler (KL) divergence [Liu and Wang, 2016], the maximum mean discrepancy [Arbel et al., 2019], or the kernelised Stein discrepancy [Fisher et al., 2021, Korba et al., 2021]. Interacting particle systems obtained in this general framework are intended to provide reasonable approximations to the posterior as algorithmic time approaches infinity ( $t \rightarrow \infty$ ) and their convergence can be analysed using the theory of gradient flows on the space of probability distributions [Ambrosio et al., 2008, Villani, 2008], see, for instance, Duncan et al. [2023], Korba et al. [2020], Chewi et al. [2020], Nüsken and Renger [2023].

In this paper, we follow an alternative construction principle, sometimes referred to as the *homotopy method* [Daum and Huang, 2008, Reich, 2011].<sup>1</sup> Rather than minimising an objective functional and relying on convergence in the long-time limit, from the outset we fix a guiding curve of probability distributions that interpolates between the prior (at  $t = 0$ ) and the posterior (at  $t = 1$ ). Such interpolations (or homotopies) are routinely used in tempering approaches to

---

<sup>1</sup>The term ‘homotopy’ is borrowed from the field of topology, where it describes deformations of one continuous function into another.

Bayesian inference, often in tandem with sequential Monte Carlo [Chopin and Papaspiliopoulos, 2020, Smith, 2013], and have recently been explored in connection with diffusion models [Vargas et al., 2024].

Similarly to Stein variational gradient descent (SVGD) [Liu and Wang, 2016], we seek a driving vector field for the particles from a reproducing kernel Hilbert space (RKHS). Based on a kernel ridge regression type formulation and the prescribed interpolation, we derive update equations that are reminiscent of those of SVGD, and that form the backbone of the proposed scheme, named *Stein transport*.

To clarify the connections between SVGD and Stein transport, we explore their shared geometric foundations rooted in a formal Riemannian structure, initially identified by Liu [2017] and further elaborated by Duncan et al. [2023], Nüsken and Renger [2023]. Specifically, while SVGD executes a gradient flow of the Kullback-Leibler (KL) divergence on the space of probability distributions – where this space is endowed with an optimal transport distance induced by the kernel – Stein transport can be viewed as implementing infinitesimally optimal transport maps within the same geometrical framework. This perspective not only justifies the name ‘Stein transport’ but also provides an alternative derivation that aligns with the broader ‘sampling via transport’ framework introduced by El Mousely and Marzouk [2012]. The fact that Stein transport can be derived from either a regression or an optimal transport perspective underscores the intrinsic connection between these two approaches: Both are centered around least-squares type objectives (see also Zhu and Mielke [2024]), a structural similarity that might become one of the key aspects in the recently emerging field of statistical optimal transport [Chewi et al., 2024].

Implementations of Stein transport are subject to numerical errors arising from both the (Tikhonov) regularisation of the regression problem and finite-particle effects. By analysing the mean-field limit of Stein transport, we examine these sources of inaccuracy and establish connections to birth-death dynamics [Lu et al., 2019b, 2023] and Fisher-Rao gradient flows [Chizat et al., 2018, Liero et al., 2018]. Based on this analysis, we propose practical guidelines to minimise these errors. Specifically, we introduce *Adjusted Stein transport*, which alternates between Stein transport and SVGD steps (where the latter are supposed to ‘adjust’ the particles, thereby increasing the accuracy of Stein transport). Our numerical experiments demonstrate that often this approach not only significantly improves on SVGD in terms of accuracy and computational cost, but also effectively addresses the variance collapse issue that is frequently encountered in SVGD [Ba et al., 2021].

**Outline.** The paper is organised as follows. In Section 2, we review the homotopy method and its relation to Stein operators. Section 3 formulates an appropriate kernel ridge regression problem, demonstrating that its unique solution can be obtained in closed form (see Proposition 2). The resulting system of equations defining Stein transport is presented in (11). Section 4 analyses the mean-field limit of Stein transport, covering both the (static) kernel ridge regression problem (Section 4.1) and the dynamics of the particle system (Section 4.2). We study in particular the impact of regularisation and finite-particle effects on the accuracy of the posterior approximation. In Section 5, we explore the geometrical aspects of Stein transport and its conceptual connections to SVGD, while Section 6 provides further connections to related work. Section 7 details the implementation of Stein transport, and – based on insights from Section 4.2 – develops the adjusted variant of Stein transport (Section 7.1). The paper concludes with numerical experiments (Section 7.2) and a discussion of conclusions and outlook (Section 8).

## 2 Interpolations in Bayesian inference and the Stein operator

For a probabilistic model  $p(x, y)$ , Bayesian inference relies on the posterior  $p(x|y) = \frac{p(y|x)p(x)}{p(y)}$ , where  $x \in \mathbb{R}^d$  is a (hidden) parameter of interest, and  $y$  represents the available data. Moreover,  $p(x)$  is the prior,  $p(y|x)$  is the likelihood, and  $p(y)$  is an intractable normalisation constant. We suppress the dependence on  $y$  and introduce the notation  $\pi_0(x) = p(x)$  for the prior, and  $\pi_1(x) = p(x|y)$  for the posterior. The likelihood will be assumed of the form  $p(y|x) = \exp(-h(x))$ , for a continuous function  $h : \mathbb{R}^d \rightarrow \mathbb{R}$ , bounded from below, and such that  $\int_{\mathbb{R}^d} \exp(-h(x))\pi_0(dx) < \infty$ .

The *homotopy method* [Daum and Huang, 2008, 2009, Daum et al., 2010, Heng et al., 2021, Reich, 2011, 2012, 2022] posits a continuous deformation in algorithmic time  $t \in [0, 1]$  that bridges the prior  $\pi_0$  and the posterior  $\pi_1$  through intermediate distributions  $\pi_t$ , defined as

$$\pi_t = \frac{e^{-th}\pi_0}{Z_t}, \quad t \in [0, 1], \tag{1}$$

where  $Z_t = \int_{\mathbb{R}^d} e^{-th} d\pi_0$  denote the respective normalisation constants. In the context of tempering and simulated annealing [Chopin and Papaspiliopoulos, 2020, Chapter 17],  $t$  plays the role of an inverse temperature parameter. Presented with samples  $X_0^1, \dots, X_0^N \in \mathbb{R}^d$  from the prior  $\pi_0$  (to be thought of as particles), our objective is to devise a dynamical scheme that moves those particles in such a way that the corresponding empirical distribution approximately

follows the interpolation (1), that is,  $\frac{1}{N} \sum_{i=1}^N \delta_{X_t^i} \approx \pi_t$ , for all  $t \in [0, 1]$ . Clearly, a successful implementation of this strategy would yield samples  $X_1^1, \dots, X_1^N$  approximately distributed according to the posterior  $\pi_1$ .

**Remark 1.** The specific form of the interpolation (1) is not essential for the developments in this paper, as long as  $\pi_0$  and  $\pi_1$  correspond to prior and posterior, respectively. It is the case, however, that the specific form of (1) can be linked to Fisher-Rao and Newton gradient flows [Chen et al., 2023b, Chopin et al., 2023, Domingo-Enrich and Pooladian, 2023, Lu et al., 2023], see Section 6.2 for a discussion. On the other hand, Syed et al. [2021] demonstrate benefits of alternative interpolations in the context of parallel tempering, and exploring these possibilities is an interesting avenue for future work.

In order to construct such schemes, we will make essential use of the *Stein operators*  $S_\pi : C^1(\mathbb{R}^d; \mathbb{R}^d) \rightarrow C(\mathbb{R}^d; \mathbb{R})$ , defined via

$$S_\pi v := \frac{1}{\pi} \nabla \cdot (\pi v) = \nabla \log \pi \cdot v + \nabla \cdot v, \quad (2)$$

which can be associated to any strictly positive and continuously differentiable probability density  $\pi$  (see Anastasiou et al. [2023] for an overview). Here and in what follows,  $\nabla \cdot v := \sum_{i=1}^d \partial_{x_i} v_i$  denotes the divergence of  $v$ . Importantly for applications in Bayesian inference, (2) can be computed given a vector field  $v$  without access to the normalisation constant of  $\pi$ : Indeed, if  $\tilde{\pi} = Z\pi$  is an unnormalised version of  $\pi$ , then  $\nabla \log \tilde{\pi} = \nabla \log \pi$ . The relevance of  $S_\pi$  for the present paper derives from the following result:

**Proposition 1** (Stein equation). *Assume that the time-dependent vector field  $v_t \in C^1(\mathbb{R}^d; \mathbb{R}^d)$  satisfies the Stein equation*

$$S_{\pi_t} v_t = h - \int_{\mathbb{R}^d} h \, d\pi_t, \quad (3)$$

for all  $t \in [0, 1]$ . Then the ordinary differential equation (ODE) with random initial condition  $\pi_0$ ,

$$\frac{dX_t}{dt} = v_t(X_t), \quad X_0 \sim \pi_0, \quad (4)$$

reproduces the interpolation (1), in the sense that  $\text{Law}(X_t) = \pi_t$ , for all  $t \in [0, 1]$ , whenever (4) is well posed.

*Proof.* See, for instance, Daum et al. [2010], Reich [2011] or Heng et al. [2021]. For the reader's convenience we provide a proof in Appendix A.2.  $\square$

**Remark 2** (Nonuniqueness). Solutions to (3) are not unique; indeed from  $S_\pi v = \frac{1}{\pi} \nabla \cdot (\pi v)$  we see that any  $\pi$ -weighted divergence-free vector field is in the kernel of  $S_\pi$ , and may hence be added to a solution of (3) without affecting its validity. Any such solution is permissible in principle, reproducing the interpolation (1) at the level of time-marginals, but leading to different trajectories (or measures on path space). In terms of (optimal) transport, the Stein equation (3) can therefore be thought of as encoding marginal constraints, without selecting any optimality principle.

Although (4) is formulated for random initial conditions  $\pi_0$ , corresponding particle movement schemes can often be obtained by approximating  $\pi_0$  with an empirical measure and applying the flow (4) to each particle separately (see (11) below). We mention in passing that the homotopy method allows us to compute the normalising constant  $Z_1 = p(y)$ , commonly used for Bayesian model selection, via the identity

$$Z_1 = \exp \left( - \int_0^1 \int_{\mathbb{R}^d} h \, d\pi_t \, dt \right),$$

see Gelman and Meng [1998], Oates et al. [2016], or the proof of Proposition 1 in Appendix A.1.

### 3 Solving the Stein equation (3) using kernel ridge regression

Proposition 1 shifts the challenge of approximating the posterior  $\pi_1$  to the problem of obtaining suitable (approximate) solutions to the Stein equation (3), given samples  $X_t^1, \dots, X_t^N \in \mathbb{R}^d$ , distributed (approximately) according to  $\pi_t$ . In this section, we present an approach based on a formulation in the spirit of kernel ridge regression [Kanagawa et al., 2018], seeking vector fields  $v_t$  in a reproducing kernel Hilbert space (RKHS)  $\mathcal{H}_k^d$ . To set the stage, we recall the relevant background (see, for example, Smola and Schölkopf [1998], Kanagawa et al. [2018] or Steinwart and Christmann [2008, Chapter 4] for more details):

**Preliminaries on positive definite kernels and RKHSs.** Throughout, we consider positive definite kernels; those are bivariate real-valued functions  $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  that are symmetric,  $k(x, y) = k(y, x)$ , for all  $x, y \in \mathbb{R}^d$ , and that satisfy the positive (semi-)definiteness condition  $\sum_{i,j=1}^N \alpha_i \alpha_j k(x_i, x_j) \geq 0$ , for all  $\alpha_1, \dots, \alpha_N \in \mathbb{R}$ ,  $x_1, \dots, x_N \in \mathbb{R}^d$ , and  $N \in \mathbb{N}$ . We will assume that  $k \in C^{1,1}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R})$ , that is, for all  $i, j = 1, \dots, d$ , the mixed partial derivatives  $\partial_{x_i} \partial_{y_j} k(x, y)$  exist and are continuous. The reproducing kernel Hilbert space (RKHS) corresponding to  $k$  is a Hilbert space of real-valued functions on  $\mathbb{R}^d$  and will be denoted by  $(\mathcal{H}_k, \langle \cdot, \cdot \rangle_{\mathcal{H}_k})$ . It is characterised by the conditions that  $k(x, \cdot) \in \mathcal{H}_k$ , for all  $x \in \mathbb{R}^d$ , as well as  $\langle f, k(x, \cdot) \rangle_{\mathcal{H}_k} = f(x)$ , for all  $x \in \mathbb{R}^d$  and  $f \in \mathcal{H}_k$  [Steinwart and Christmann, 2008, Section 4.2]. The  $d$ -fold Cartesian product

$$\mathcal{H}_k^d = \underbrace{\mathcal{H}_k \times \dots \times \mathcal{H}_k}_{d \text{ times}}$$

consists of vector fields  $v = (v_1, \dots, v_d)$  with component functions  $v_i \in \mathcal{H}_k$  and is equipped with the inner product  $\langle u, v \rangle_{\mathcal{H}_k^d} := \sum_{i=1}^d \langle u_i, v_i \rangle_{\mathcal{H}_k}$ .

We are now in a position to present our kernel-based approach towards approximating solutions to the Stein equation (3). The following formulation seeks to minimise the difference between the left- and right-hand sides of (3), based on available samples:

**Problem 1** (Kernel ridge regression for the Stein equation (3)). *Given samples  $X^1, \dots, X^N \in \mathbb{R}^d$ , a strictly positive probability density  $\pi \in C^1(\mathbb{R}^d; \mathbb{R}_{>0})$ , and a regularisation parameter  $\lambda > 0$ , find a solution to*

$$v^* \in \operatorname{argmin}_{v \in \mathcal{H}_k^d} \left( \frac{1}{N} \sum_{j=1}^N ((S_\pi v)(X^j) - h_0(X^j))^2 + \lambda \|v\|_{\mathcal{H}_k^d}^2 \right), \quad (5)$$

where  $h_0$  refers to the data-centred version of  $h$ , that is,  $h_0(x) := h(x) - \frac{1}{N} \sum_{j=1}^N h(X^j)$ .

The regularising term  $\lambda \|v\|_{\mathcal{H}_k^d}^2$  guarantees strict convexity of the objective in (5) and will turn out to stabilise the associated numerical procedure (in particular, the inversion of a linear system). In our experiments, we typically choose  $\lambda$  to be small, say  $\lambda \approx 10^{-3}$ . Larger values of  $\lambda$  may be appropriate when the likelihood itself is noisy or a stronger form of regularisation appears to be suitable for other reasons (see, for instance, Dunbar et al. [2022]). The formulation in Problem 1 is classical if  $S_\pi$  is replaced by the identity operator, the task in this case being to recover an underlying target function from noisy measurements [Kanagawa et al., 2018, Section 3.2]. We would also like to point the reader to the work of Zhu and Mielke [2024], where regression type formulations are used to construct dynamical schemes for inference (but the specifics of our formulation are quite different). In the context of PDEs, this approach is often linked to collocation methods, see [Wendland, 2004, Chapter 16], Fasshauer [2007, Chapter 38]; for a recent nonlinear extension, see Chen et al. [2021].

**The KSD-kernel.** In order order to address Problem 1, recall the following from the theory of *kernelised Stein discrepancies* (KSD) [Liu et al., 2016, Gorham and Mackey, 2017]:

Given a positive definite kernel  $k$  and a score function  $\nabla \log \pi$ , we can construct a new positive definite kernel  $\xi^{k, \nabla \log \pi} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ , defined as

$$\begin{aligned} \xi^{k, \nabla \log \pi}(x, y) := S_\pi^x S_\pi^y k(x, y) &= \nabla \log \pi(x) \cdot \nabla_y k(x, y) + \nabla \log \pi(y) \cdot \nabla_x k(x, y) \\ &\quad + \nabla_x \cdot \nabla_y k(x, y) + \nabla \log \pi(x) \cdot k(x, y) \nabla \log \pi(y), \end{aligned} \quad (6)$$

in the following referred to as the *KSD-kernel* associated to  $k$  and  $\nabla \log \pi$ . In (6) we have used the notation  $\nabla_x \cdot \nabla_y k(x, y) := \sum_{i=1}^d \partial_{x_i} \partial_{y_i} k(x_i, y_i)$ , and  $S_\pi^x$  (respectively  $S_\pi^y$ ) is understood to act on the variable  $x$  (respectively  $y$ ) only. The salient feature of  $\xi^{k, \nabla \log \pi}$  is that it integrates to zero against  $\pi$ ,

$$\int_{\mathbb{R}^d} \xi^{k, \nabla \log \pi}(\cdot, y) \pi(dy) = 0,$$

and that, as a consequence, the kernelised Stein discrepancy

$$\text{KSD}(\mu | \pi) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \xi^{k, \nabla \log \pi}(x, y) \mu(dx) \mu(dy) \quad (7)$$

can be used to measure similarity between the two probability measures  $\mu$  and  $\pi$  [Liu et al., 2016, Theorem 3.6].

The minimisation in (5) admits a unique solution that can now be represented in closed form: We define the Gram matrix  $\xi^{k,\nabla \log \pi} \in \mathbb{R}^{N \times N}$  and the vector  $\mathbf{h}_0 \in \mathbb{R}^N$  obtained from evaluating the KSD-kernel  $\xi^{k,\nabla \log \pi}$  and the centred negative log-likelihood  $h_0$  (defined in Problem 1) at the locations of the particles,

$$(\xi^{k,\nabla \log \pi})_{ij} = \xi^{k,\nabla \log \pi}(X^i, X^j) \in \mathbb{R}^{N \times N}, \quad (\mathbf{h}_0)_i = h(X^i) \in \mathbb{R}^N. \quad (8)$$

Furthermore, we denote by  $I_{N \times N} \in \mathbb{R}^{N \times N}$  the identity matrix of dimension  $N$ . Using this notation, we have the following result.

**Proposition 2.** *For  $\lambda > 0$ , the kernel ridge regression problem (5) admits a unique solution, given by*

$$v^* = \frac{1}{N} \sum_{j=1}^N \phi^j (k(\cdot, X^j) \nabla \log \pi(X^j) + \nabla_{X^j} k(\cdot, X^j)), \quad (9)$$

where  $(\phi^j)_{j=1}^N = \phi \in \mathbb{R}^N$  is the unique solution to the linear system

$$\left( \frac{1}{N} \xi^{k,\nabla \log \pi} + \lambda I_{N \times N} \right) \phi = \mathbf{h}_0. \quad (10)$$

**Remark 3.** Since  $\xi^{k,\nabla \log \pi}$  is a positive definite kernel, the matrix  $\xi^{k,\nabla \log \pi} + \lambda I_{N \times N}$  is invertible, and hence (10) determines  $\phi$  uniquely.

*Proof.* There are (at least) two ways of arriving at the representation of  $v^*$  given by (9) and (10): Firstly, we may rely on (a version of) the representer theorem [Wahba and Wang, 2019], leveraging the fact that for each  $j = 1, \dots, N$ , the mapping  $v \mapsto (S_\pi v)(X^j)$  is a continuous linear functional on  $\mathcal{H}_k^d$ , and that the corresponding Riesz representers are given by  $k(\cdot, X^j) \nabla \log \pi(X^j) + \nabla_{X^j} k(\cdot, X^j)$ . Alternatively, we can think of (5) as a Tikhonov-regularised least squares problem [Kirsch, 2021, Section 2.2], whose solution  $S_{\pi,N}^*(\lambda I_{N \times N} + S_{\pi,N} S_{\pi,N}^*)^{-1} \mathbf{h}_0$  can be identified with  $v^*$ .<sup>2</sup> We detail both approaches in Appendix A.1.  $\square$

Combining Propositions 1 and 2, we obtain the following interacting particle system, defining Stein transport:

$$\begin{cases} \frac{dX_t^i}{dt} = \frac{1}{N} \sum_{j=1}^N \phi_t^j \left( k(X_t^i, X_t^j) \nabla \log \pi_t(X_t^j) + \nabla_{X_t^j} k(X_t^i, X_t^j) \right), & t \in [0, 1], \\ \left( \frac{1}{N} \xi^{k,\nabla \log \pi_t} + \lambda I_{N \times N} \right) \phi_t = \mathbf{h}_{0,t}. \end{cases} \quad (11)$$

Note that for  $\xi^{k,\nabla \log \pi_t}$  and the update  $\frac{dX_t^i}{dt}$  in (11), the score function  $\nabla \log \pi_t$  is assumed to be calculated according to the interpolation (1), that is,  $\nabla \log \pi_t = -t \nabla h + \nabla \log \pi_0$ , where  $\nabla \log \pi_0$  is typically tractable as the score function associated to the prior (but see Section 6.1 for remarks on the score-free setting). The vector  $(\mathbf{h}_{0,t})_i = h(X_t^i) - \frac{1}{N} \sum_{j=1}^N h(X_t^j)$  is time-dependent through evaluations at the particles' locations. A standard Euler discretisation of (11) now yields an implementable procedure for Bayesian inference that we summarise in Algorithm 1.

**Comparison to SVGD.** The system (11) bears a remarkable similarity to

$$\frac{dX_t^i}{dt} = \frac{1}{N} \sum_{j=1}^N \left( k(X_t^i, X_t^j) \nabla \log \pi(X_t^j) + \nabla_{X_t^j} k(X_t^i, X_t^j) \right), \quad t \in [0, \infty), \quad (12)$$

the governing equations of Stein variational gradient descent (SVGD) introduced by Liu and Wang [2016]. Postponing a more conceptual (geometric) comparison to Section 5, we note that (12) involves the score function  $\nabla \log \pi$  associated to the target (that is,  $\nabla \log \pi_1$  in the notation of this paper), whereas (11) is driven by the time-dependent score function  $\nabla \log \pi_t$  induced by the interpolation (1). Stein transport furthermore relies on the vector  $\phi_t \in \mathbb{R}^N$ , the components of which can be interpreted as weights attached to the particles (we will provide a more in-depth discussion of the role of  $\phi$  in Section 4). In terms of computational cost, Stein transport requires the assembly and inversion of the  $N$ -dimensional linear system  $(\frac{1}{N} \xi + \lambda I_{N \times N}) \phi = \mathbf{h}$ , where  $N$  is the number of particles. We would like to stress, however, that Stein transport and SVGD demand the same number of gradient evaluations of the log-likelihood per time step, and that for moderately sized particle systems (say  $N \approx 10^3$ ), the inversion of a linear system only incurs a very minor computational overhead. For large-scale applications, speed-ups may be achieved using random feature expansions of  $k$  [Rahimi et al., 2007] or projections combined with preconditioning [Rudi et al., 2017].

Despite these similarities, Stein transport and SVGD exhibit important differences that affect their numerical performance. First, Stein transport achieves its posterior approximation at  $t = 1$ , whereas SVGD relies on long-time convergence towards a fixed point in (12). As demonstrated experimentally in Section 7, the finite-time property of Stein transport often significantly reduces the number of necessary time steps to achieve a prescribed accuracy. Perhaps more importantly, Stein transport does not suffer from particle collapse in high dimensions as does SVGD [Ba et al., 2021]. This is further elaborated in Proposition 5, Remark 9 and Section 7.2.2 below.

<sup>2</sup>The sample versions  $S_{\pi,N}$  of the Stein operator  $S_\pi$  will be introduced in Appendix A.1.

**Algorithm 1** Stein transport

---

**Input:** Prior samples  $X_0^1, \dots, X_0^N \in \mathbb{R}^d$ , prior score function  $\nabla \log \pi_0 : \mathbb{R}^d \rightarrow \mathbb{R}^d$ , negative log-likelihood  $h : \mathbb{R}^d \rightarrow \mathbb{R}$ , with gradient  $\nabla h : \mathbb{R}^d \rightarrow \mathbb{R}^d$ , positive definite kernel  $k$ , regularisation parameter  $\lambda > 0$ , time discretisation  $0 = t_0 < t_1 < \dots < t_{N_{\text{steps}}} = 1$  with time step  $\Delta t$ .

**for**  $n = 0, \dots, N_{\text{steps}} - 1$  **do**

- Compute the scores  $P_n^j = -t_n \nabla h(X_n^j) + \nabla \log \pi_0(X_n^j)$ .
- Compute the KSD Gram matrix

$$\boldsymbol{\xi}_{ij} = P_n^i \cdot \nabla_{X_n^j} k(X_n^i, X_n^j) + P_n^j \cdot \nabla_{X_n^i} k(X_n^i, X_n^j) \\ + \nabla_{X_n^i} \cdot \nabla_{X_n^j} k(X_n^i, X_n^j) + P_n^i \cdot k(X_n^i, X_n^j) P_n^j.$$

- Compute the centred negative log-likelihoods  $\mathbf{h}_i = h(X_n^i) - \frac{1}{N} \sum_{i=1}^N h(X_n^i)$ .
- Solve  $(\frac{1}{N} \boldsymbol{\xi} + \lambda I_{N \times N})\phi = \mathbf{h}$  for  $\phi \in \mathbb{R}^N$ .
- $X_{n+1}^i \leftarrow X_n^i + \frac{\Delta t}{N} \sum_{j=1}^N \phi^j \left( k(X_n^i, X_n^j) P_n^j + \nabla_{X_n^i} k(X_n^i, X_n^j) \right)$ .

**end for**

**return** approximate posterior samples  $X_{N_{\text{steps}}}^1, \dots, X_{N_{\text{steps}}}^N$ .

---

**Stein transport without gradients.** We mention in passing that Stein transport can be implemented without gradient evaluations of the log-likelihood. Indeed, the time-dependent score may be evolved along the particle flow, starting from  $\nabla \log \pi_0$ :

**Lemma 1** (Evolution of the score function). *Let  $v_t \in C^2(\mathbb{R}^d; \mathbb{R}^d)$  be a family of vector fields such that*

$$\frac{dX_t}{dt} = v_t(X_t), \quad X_0 \sim \pi_0,$$

*is well posed, and denote the corresponding laws by  $\pi_t = \text{Law}(X_t)$ . Assume for all  $t \geq 0$  that  $\pi_t \in C^1(\mathbb{R}^d; \mathbb{R})$  with  $\pi_t > 0$  and define  $P_t := \nabla \log \pi_t(X_t)$ . Then it holds that*

$$\frac{dP_t}{dt} = -\nabla(\nabla \cdot v_t)(X_t) - (\nabla v_t)(X_t)P_t.$$

*Proof.* This follows by direct calculation, see Appendix A.2.  $\square$

Based on (9), the terms  $\nabla(\nabla \cdot v^*)$  and  $\nabla v^*$  can be computed in closed form, applying the differential operators to the kernel  $k$ . The corresponding gradient-free interacting particle system is thus given by

$$\begin{cases} \frac{dX_t^i}{dt} = \frac{1}{N} \sum_{j=1}^N \phi_t^j \left( k(X_t^i, X_t^j) P_t^j + \nabla_{X_t^j} k(X_t^i, X_t^j) \right), & t \in [0, 1], \\ \frac{dP_t^i}{dt} = -\frac{1}{N} \sum_{j=1}^N \phi_t^j \left( \nabla_{X_t^i}^2 k(X_t^i, X_t^j) P_t^j + \nabla_{X_t^i} \left( \nabla_{X_t^j} k(X_t^i, X_t^j) \right) \right), \\ (\frac{1}{N} \boldsymbol{\xi}^{k, \nabla \log \pi_t} + \lambda I_{N \times N})\phi_t = \mathbf{h}_{0,t}, \end{cases} \quad (13)$$

where the initial samples  $(X_0^i)_{i=1}^N$  are taken from the prior  $\pi_0$ , and the scores are initialised as  $P_0^i = \nabla \log \pi_0(X_0^i)$ . Using (13), Algorithm 1 can straightforwardly be extended so that it does not require gradient evaluations of  $h$ .

## 4 Convergence and mean-field description

In this section, we investigate the infinite-particle limit of Stein transport, deriving formulations based on the limiting measures  $\pi_t = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \delta_{X_t^i}$ . In Section 4.1, we first examine the infinite-particle version of the kernel ridge regression Problem 1. Building on this, we draw qualitative and quantitative conclusions about the dynamics (11) and its mean-field version in Section 4.2, particularly discussing the errors incurred by regularisation ( $\lambda > 0$ ) and finite-particle approximation ( $N < \infty$ ).

#### 4.1 Infinite-particle kernel ridge regression

Formally taking the limit  $\frac{1}{N} \sum_{i=1}^N \delta_{X^i} \xrightarrow{N \rightarrow \infty} \pi$  in (5), we arrive at the following mean-field version of the kernel ridge regression task stated as Problem 1:

**Problem 2** (Kernel ridge regression for the Stein equation (3), mean-field version). *Given a strictly positive probability density  $\pi \in C^1(\mathbb{R}^d; \mathbb{R})$  and a regularisation parameter  $\lambda > 0$ , find a solution to*

$$v^* \in \operatorname{argmin}_{v \in \mathcal{H}_k^d} \left( \int_{\mathbb{R}^d} ((S_\pi v) - h_{0,\infty})^2 d\pi + \lambda \|v\|_{\mathcal{H}_k^d}^2 \right), \quad (14)$$

where  $h_{0,\infty}$  refers to the  $\pi$ -centred version of  $h$ , that is,  $h_{0,\infty}(x) := h(x) - \int_{\mathbb{R}^d} h d\pi$ .

Notice that Problem 2 is in fact a generalisation (and not only a mean-field version) of Problem 1: Choosing  $\pi = \frac{1}{N} \sum_{i=1}^N \delta_{X^i}$  in (14) leads back to (5), and in fact both formulations could be treated on an equal footing, as done by Steinwart and Christmann [2008, Chapter 5] for conventional kernel ridge regression. To address Problem 2, in the remainder of this section we will assume the following:

**Assumption 1.** *The probability density  $\pi \in C^1(\mathbb{R}^d; \mathbb{R})$  is strictly positive, and such that  $\|h\|_{L^2(\pi)} < \infty$  as well as  $\|\nabla \log \pi\|_{(L^2(\pi))^d} < \infty$ . The kernel  $k$  is bounded, with bounded first-order partial derivatives.*

The solution to Problem 2 will be given in terms of the integral operator

$$\mathcal{T}_{k,\pi}\phi = \int_{\mathbb{R}^d} k(\cdot, y)\phi(y)\pi(dy), \quad \phi \in L^2(\pi), \quad (15)$$

and we refer to Steinwart and Christmann [2008, Chapter 4.3] for some of its properties. Interpreting (15) component-wise, we can also apply  $\mathcal{T}_{k,\pi}$  to vector-valued functions, that is, when  $\phi \in (L^2(\pi))^d$ . In parallel to Proposition 2, we can now characterise the unique solution to Problem 2:

**Proposition 3.** *Assume that  $\pi$ ,  $h$  and  $k$  satisfy Assumption 1. Then there exists a unique solution to (14), given by  $v^* = -\mathcal{T}_{k,\pi}\nabla\phi$ , where  $\phi \in L^2(\pi)$  is the unique solution to the Stein-Poisson equation*

$$-\nabla \cdot (\pi \mathcal{T}_{k,\pi} \nabla \phi) + \lambda \pi \phi = \pi \left( h - \int_{\mathbb{R}^d} h d\pi \right). \quad (16)$$

*Proof.* As in the proof of Proposition 2, we can interpret (14) as a Tikhonov-regularised least-squares problem. Details can be found in Appendix B.  $\square$

The Stein-Poisson equation (16) is a kernelised and regularised version of the (conventional) Poisson equation

$$-\nabla \cdot (\pi \nabla \phi) = \pi \left( h - \int_{\mathbb{R}^d} h d\pi \right), \quad (17)$$

which governs central limit theorems and fluctuations in Markov processes [Komorowski et al., 2012, Pardoux and Veretennikov, 2001] and also features prominently in McKean-Vlasov approaches to nonlinear filtering [Laugesen et al., 2015, Pathiraja et al., 2021, Coghi et al., 2023] as well as in the development of particle-flow algorithms on the basis of Proposition 1 [Heng et al., 2021, Reich, 2011, Taghvaei and Mehta, 2016, Radhakrishnan and Meyn, 2019, Taghvaei et al., 2020, Maurais and Marzouk, 2023, 2024, Wang and Nüsken, 2024, Taghvaei et al., 2020, Tian et al., 2024]. We mention in passing that the Stein-Poisson equation (16) has implicitly been used by Oates et al. [2017] to construct control variates for Monte Carlo estimators.

The following two remarks compare (16) and (17) in terms of well-posedness and finite-particle approximations: In a nutshell, the Stein-Poisson equation (16) is inferior in terms of well-posedness, but superior in terms of finite-particle approximations. We will comment on the geometrical underpinnings of (16) and (17) in Section 5.

**Remark 4** (Well-posedness). Although at first glance the expression on the left-hand side of (16) is only well defined if  $\phi$  is differentiable, the operator  $\phi \mapsto \nabla \cdot (\pi \mathcal{T}_{k,\pi} \nabla \phi)$  can be extended to a bounded linear operator on  $L^2(\pi)$ , see Lemma 6 in Appendix B. Also note that  $\lambda > 0$  is necessary (and sufficient) for (16) to be well posed. Indeed, since the operator  $\phi \mapsto -\nabla \cdot (\pi \mathcal{T}_{k,\pi} \nabla \phi)$  is compact in  $L^2(\pi)$  according to Duncan et al. [2023, proof of Lemma 23], the Stein-Poisson equation (16) is ill posed for  $\lambda = 0$ , in the sense of Kirsch [2021, Section 1.2]: A solution  $\phi$  will in general not exist, and, even if it does, it will necessarily be unstable with respect to small perturbations of  $h$ . In contrast, the Poisson equation (17) is well posed without regularisation, under mild conditions on the tails of  $\pi$  (see, for instance, Lelievre and Stoltz [2016, Corollary 2.4]).

**Remark 5** (Approximation by empirical measures). In the same vein as (14), the Stein-Poisson equation (16) is in fact meaningful for  $\pi = \frac{1}{N} \sum_{i=1}^N \delta_{X^i}$ . Indeed, dividing (16) by  $\pi$  and manipulating the left-hand side, we arrive at

$$\int_{\mathbb{R}^d} \xi^{k, \nabla \log \pi_t}(\cdot, y) \phi(y) \pi(dy) + \lambda \phi = h - \int_{\mathbb{R}^d} h d\pi,$$

which makes sense for  $\pi = \frac{1}{N} \sum_{i=1}^N \delta_{X^i}$ , and reduces to the linear system (10) under this substitution (note that we need to keep  $\nabla \log \pi_t$  as is, as an ‘exterior input’). In contrast, there is no direct interpretation for  $\pi = \frac{1}{N} \sum_{i=1}^N \delta_{X^i}$  of the conventional Poisson equation (17), and additional approximations (or finite dimensional projections) are necessary to obtain a particle-based formulation. This difficulty can be traced back to the following observation: The Poisson equation (17) formally describes minimisers in

$$v^* \in \operatorname{argmin}_{v \in L^2(\pi)} \left( \int_{\mathbb{R}^d} ((S_\pi v) - h_0)^2 d\pi + \lambda \|v\|_{L^2(\pi)}^2 \right),$$

for  $\lambda \rightarrow 0$ , replacing  $\|\cdot\|_{\mathcal{H}_k^d}$  in (14) by  $\|\cdot\|_{(L^2(\pi))^d}$ . Unfortunately, this replacement cannot be done in (5), since  $S_\pi v(X^i)$  would be ill defined (point evaluations are not possible for  $v \in (L^2(\pi))^d$ ).

We end this subsection by establishing rigorous correspondences between the finite-sample kernel ridge regression in Problem 1 and its mean-field version in Problem 2. Under appropriate growth conditions on  $\nabla \log \pi$  and  $h$ , the corresponding optimal vector fields converge in the  $\mathcal{H}_k^d$ -norm:

**Theorem 1** (Connections between Problems 1 and 2). *Let  $\pi, h$  and  $k$  satisfy Assumption 1. Furthermore, assume that there exist constants  $C > 0$  and  $p > 1$  such that  $\nabla \log \pi$  and  $h$  satisfy the growth conditions*

$$|\nabla \log \pi(x)| \leq C(1 + |x|^{p/2}) \quad \text{and} \quad |h(x)| \leq C(1 + |x|^{p/2}), \quad \text{for all } x \in \mathbb{R}^d.$$

Let  $(X^i)_{i=1}^\infty \subset \mathbb{R}^d$  be a sequence of points such that the empirical measures  $\frac{1}{N} \sum_{i=1}^N \delta_{X^i}$  converge to  $\pi$  in the  $p$ -Wasserstein distance, as  $N \rightarrow \infty$ .<sup>3</sup> Denote by  $v_{N,\lambda}^*$  and  $v_{\infty,\lambda}^*$  the minimisers of (5) and (14), respectively, and fix  $\lambda > 0$ . Then  $v_{N,\lambda}^*$  converges to  $v_{\infty,\lambda}^*$  in  $\mathcal{H}_k^d$  as  $N \rightarrow \infty$ , that is,

$$\|v_{N,\lambda}^* - v_{\infty,\lambda}^*\|_{\mathcal{H}_k^d} \xrightarrow{N \rightarrow \infty} 0.$$

*Proof.* See Appendix B. The proof adapts techniques and results from statistical learning theory [Smale and Zhou, 2007, Blanchard and Mücke, 2018] to the setting of Problems 1 and 2.  $\square$

**Remark 6.** Under Assumption 1, convergence in  $\mathcal{H}_k^d$  implies uniform convergence of  $v_{N,\lambda}^*$  and its first derivatives, see Steinwart and Christmann [2008, Lemmas 4.23 and 4.34].

Our next result establishes the sense in which the solution to Problem 1 (given in Proposition 2) provides an approximate solution to the Stein equation (3). Clearly, in order to obtain satisfactory guarantees, the ‘search space’  $\mathcal{H}_k^d$  has to be sufficiently expressive. We formalise this intuition as follows:

**Assumption 2** (Universality). *The inclusion  $\mathcal{H}_k^d \subset (L^2(\pi))^d$  is dense.*

The Gaussian, Laplacian, inverse multiquadratic and Matérn kernels are universal in the sense of Assumption 2, under mild conditions on  $\pi$ , and we refer the reader to Sriperumbudur et al. [2011] for comprehensive results and a detailed overview. We can now state the following result, showing that indeed Proposition 2 yields a good approximate solution to the Stein equation (3), provided that  $\lambda$  is small and  $N$  is large enough.

**Theorem 2.** *Assume the setting from Theorem 1, and furthermore that Assumption 2 is satisfied. Then, for all  $\varepsilon > 0$  there exists  $\lambda > 0$  and  $N_0 \in \mathbb{N}$  such that*

$$\left\| S_\pi v_{N,\lambda}^* - \left( h - \int_{\mathbb{R}^d} h d\pi \right) \right\|_{L^2(\pi)} < \varepsilon, \tag{18}$$

for all  $N > N_0$ .

*Proof.* See Appendix B.  $\square$

---

<sup>3</sup>Note that this implicitly implies that the  $p^{\text{th}}$  moment of  $\pi$  is finite.

**Remark 7.** The statement of Theorem 2 requires  $\lambda$  to tend to zero ‘more slowly than  $N$  tends to infinity’. Indeed, this is a necessary requirement as the Stein-Poisson equation (16) describing the  $N \rightarrow \infty$  limit becomes ill posed for  $\lambda = 0$ , and is typical of regularisation strategies for inverse problems [Kirsch, 2021].

**Remark 8** (Source conditions for  $h$ ). In order to estimate the posterior approximation error incurred by Stein transport, it would be highly desirable to obtain versions of (18) controlling stronger norms than  $\|\cdot\|_{L^2(\pi)}$ , with more quantitative convergence rates in terms of  $N$ ,  $\varepsilon$  and  $\lambda$ . For this, additional assumptions on  $h$  would be required, for instance source conditions [Engl et al., 1996, Section 3.2] of the form

$$h - \int_{\mathbb{R}^d} h \, d\pi \in \mathcal{H}_{\xi^{k,\nabla \log \pi}}^\alpha, \quad 0 < \alpha \leq 1, \quad (19)$$

requiring the centred negative log-likelihood to belong to a fractional power of the RKHS [Muandet et al., 2017, Definition 4.11] associated to the KSD-kernel  $\xi^{k,\nabla \log \pi}$ . The spaces  $\mathcal{H}_{\xi^{k,\nabla \log \pi}}^\alpha$  interpolate between  $L^2(\pi)$  and  $\mathcal{H}_{\xi^{k,\nabla \log \pi}}$ , and thus (19) should be interpreted as an additional (abstract) smoothness assumption: The negative log-likelihood  $h$  and  $\xi^{k,\nabla \log \pi}$  are more aligned (or  $h$  is more regular when regularity is measured in terms of  $\xi^{k,\nabla \log \pi}$ ) if (19) holds with greater powers of  $\alpha$ . Fine analysis of (19) could potentially inform the choice of  $k$  and is left for future work; an application of (19) will be presented below in Proposition 4.

## 4.2 Mean field dynamics

According to Theorem 2, the driving vector field constructed from the formulation in Problem 1 satisfies the Stein equation (3) in an approximate sense. In this section, we discuss the impact of regularisation ( $\lambda > 0$ ) and finite-particle effects ( $N < \infty$ ) on the dynamics of the particle system, and in particular on the posterior approximation at final time  $t = 1$ .

### 4.2.1 Impact of the regularisation ( $\lambda > 0$ ): weighted/birth-death dynamics

If  $\phi$  solves the Stein-Poisson equation (16), it follows by integrating both sides that  $\int \phi \, d\pi = 0$ . Therefore, the regularisation has the same effect as a modification of the negative log-likelihood: We can bring  $\lambda \pi \phi$  to the right-hand side and redefine  $\tilde{h}_t := h - \lambda \phi_t$  to absorb the regularising term. As a consequence, we expect that, asymptotically as  $N \rightarrow \infty$ , Stein transport solves a slightly different Bayesian inference problem, associated to  $\tilde{h}_t$ . The following corollary to the proof of Theorem 2 shows that if  $\lambda$  is small, then  $\tilde{h}_t$  is close to  $h$  (and thus we expect to recover the original inference problem as  $\lambda \rightarrow 0$ ):

**Proposition 4.** *Let Assumptions 1 and 2 be satisfied, and denote by  $\phi^{(\lambda)} \in L^2(\pi)$  the solution to the Stein-Poisson equation (16), with regularisation parameter  $\lambda > 0$ . Then  $\lambda \phi^{(\lambda)}$  converges to zero in  $L^2(\pi)$  as  $\lambda \rightarrow 0$ . If, moreover,  $h$  satisfies the source condition (19) for some  $\alpha \in (0, 1]$ , then there exists a constant  $C_\alpha > 0$  such that*

$$\|\lambda \phi^{(\lambda)}\|_{L^2(\pi)} \leq C_\alpha \lambda^\alpha, \quad \lambda > 0. \quad (20)$$

*Proof.* See Appendix B.  $\square$

The estimate (20) shows that it is desirable to align  $h$  with  $\xi^{k,\nabla \log \pi}$  in the sense of (19), as the correction  $\lambda \phi^{(\lambda)}$  vanishes more quickly for larger values of  $\alpha$  (cf. Remark 8). Proposition 4 gives a theoretical handle on the impact of regularisation, but the ‘modified likelihood interpretation’ of the regularising term  $\lambda \pi \phi$  in (16) also suggests a correction scheme to debias Stein transport: we can attach weights  $(w_t^i)_{i=1}^N$  to the particles  $(X_t^i)_{i=1}^N$  that absorb the error term  $-\lambda \phi_t^i$ . More specifically, the Stein-Poisson equation (16) allows us to write

$$\underbrace{\nabla \cdot (\mathcal{T}_{k,\pi_t} \nabla \phi_t)}_{\text{transport}} \underbrace{-\lambda \phi_t \pi_t}_{\text{reweighting}} = -\pi_t \left( h - \int_{\mathbb{R}^d} h \, d\pi_t \right) = \partial_t \pi_t, \quad (21)$$

where the second equality follows if we assume that  $(\pi_t)_{t \in [0,1]}$  satisfies the interpolation (1), see (51) in Appendix A.2. It is instructive to perform integration by parts in  $\nabla \cdot (\mathcal{T}_{k,\pi_t} \nabla \phi_t)$ , rewriting (21) as

$$\partial_t \pi_t = -\nabla \cdot \left( \pi_t \int_{\mathbb{R}^d} (k(\cdot, y) \nabla \log \pi_t(y) + \nabla_y k(x, y)) \phi_t(y) \pi_t(dy) \right) - \lambda \phi_t \pi_t. \quad (22)$$

The coupled system comprised of (16) and (22) are the mean-field equations of (weighted) Stein transport, and, as expected, they are similar to the SVGD mean field limit [Lu et al., 2019a, Duncan et al., 2023]. Equations (21) and (22)

motivate the *weighted interacting particle system*

$$\frac{dX_t^i}{dt} = v_t(X_t^i) \quad (23a)$$

$$\frac{dw_t^i}{dt} = -\lambda w_t^i \phi_t^i, \quad w_0^i = \frac{1}{N}, \quad i = 1, \dots, N, \quad (23b)$$

where  $(\phi_t^i)_{i=1}^N$  and  $v_t$  are determined from  $(X_t^i)_{i=1}^N$  as in Proposition 2 (or, from the weighted generalisation of Problem 1 discussed in Remark 11 in Appendix A.1). The combination of transporting and reweighting the particles is similar in spirit to the dynamical schemes proposed by Lu et al. [2019b, 2023], Yan et al. [2023], Gladin et al. [2024].

The (weighted) empirical measures corresponding to (23),

$$\rho_t^{(N)}(f) := \sum_{i=1}^N w_t^i f(X_t^i), \quad f \in C_b(\mathbb{R}^d), \quad (24)$$

should recover the interpolation (1) as  $N \rightarrow \infty$ , for any value of  $\lambda > 0$ , as the mean field limit is formally given by (21), which is (by construction) satisfied by the interpolation (1). However, a rigorous proof under realistic assumptions appears to be difficult, as the weights in (23b) might degenerate as  $N \rightarrow \infty$  and therefore the weighted empirical measure (24) might not have a limit. Indeed from a practical perspective, resampling schemes are typically required to prevent weight degeneracy for dynamical schemes involving weights [Del Moral et al., 2006, Chopin and Papaspiliopoulos, 2020]. It is an interesting direction to amend Algorithm 1 on the basis of (23) and to include appropriate reweighting; in this paper, we prefer to choose  $\lambda$  small enough so that the weights in (23b) remain almost constant (as suggested by Proposition 4) retaining a purely transport-based scheme. Ignoring the weight update (23b) incurs a small asymptotic bias in the posterior approximation, but forcing equal weights is preferable in terms of effective sample size [Chopin and Papaspiliopoulos, 2020, Section 8.6]. We also stress (again) that aligning  $h$  and  $\xi^{k,\nabla \log \pi}$  in the sense of (19) is expected to be beneficial according to Proposition 4: Smaller values of  $\lambda \phi_t^i$  in (23b) lead to smaller weight updates (incurring a smaller asymptotic bias or, for the weighted scheme, a larger effective sample size).

#### 4.2.2 Impact of the finite particle approximation ( $N < \infty$ ): projection from the mean-field limit

The following proposition gives some insight into the nature of the error induced by the finite-particle approximation. As in Theorem 1, we denote by  $v_{N,\lambda}^*$  the solution to Problem 1.

**Proposition 5** (Projection). *Assume that  $h$  and  $\pi$  are such that there exists a solution  $v = -\mathcal{T}_{k,\pi} \nabla \phi \in \mathcal{H}_k^d$  to the Stein equation (3); that is,  $\phi$  solves the Stein-Poisson equation (16) for  $\lambda = 0$ . Then, for any selection of points  $\mathbf{X} = (X^1, \dots, X^N) \in (\mathbb{R}^d)^N$  such that the Gram matrix  $\xi^{k,\nabla \log \pi} \in \mathbb{R}^{N \times N}$  is invertible and  $\frac{1}{N} \sum_{i=1}^N h(X^i) = \int_{\mathbb{R}^d} h d\pi$ , there exist an orthogonal projection  $P_{\mathbf{X}}$  on  $\mathcal{H}_k^d$  such that  $v_{N,0}^* = P_{\mathbf{X}} v$ . In particular,  $\|v_{N,0}^*\|_{\mathcal{H}_k^d} \leq \|v\|_{\mathcal{H}_k^d}$ .*

*Proof.* See Appendix B. The orthogonal projection  $P_{\mathbf{X}}$  projects onto the subspace

$$\text{span} \{ \nabla_{X^i} k(\cdot, X^i) + k(\cdot, X^i) \nabla \log \pi(X^i) : i = 1, \dots, N \} \subset \mathcal{H}_k^d, \quad (25)$$

spanned by the ‘SVGD vector fields’ at the particle positions.  $\square$

The conditions in Proposition 5 are very restrictive (existence of a solution in the relatively small space  $\mathcal{H}_k^d$ , see Remark 4, and  $\frac{1}{N} \sum_{i=1}^N h(X^i) = \int_{\mathbb{R}^d} h d\pi$ ), but a more comprehensive and quantitative version could be obtained with more technical effort. The main message from Proposition 5 is that the vector field (9) based on  $N$  particles tends to underestimate  $v_{\infty}^* = -\mathcal{T}_{k,\pi} \nabla \phi$  obtained from the mean-field formulation in Proposition 3 (if this comparison is made in the  $\mathcal{H}_k^d$ -norm). Consequently, it is reasonable to expect that Algorithm 1 tends to move the particles too slowly, and that the posterior spread is hence overestimated (as the prior is typically more spread out than the posterior). Moreover, Stein transport is expected to be more accurate in cases where the projection  $P_{\mathbf{X}}$  is close to the identity operator, that is, when the linear span in (25) nearly exhausts  $\mathcal{H}_k^d$ . Intuitively, this reasoning suggests that the particles  $(X^i)_{i=1}^N$  should be spread out (‘adjusted’) as much as possible, and accordingly we propose a modification (‘adjusted Stein transport’) in Section 7.1.

**Remark 9** (SVGD underestimates the posterior spread). In contrast to Stein transport, SVGD tends to return posterior approximations that are too peaked in comparison to the true posterior [Ba et al., 2021]. This observation can be understood from the fact that SVGD by construction minimises the reverse KL-divergence and therefore intrinsically

suffers from mode collapse [Blei et al., 2017]. As an alternative explanation, note that SVGD is a ‘noiseless’ version of the interacting diffusion [Gallego and Insua, 2018]

$$dX_t^i = \frac{1}{N} \sum_{j=1}^N \left( -k(X_t^i, X_t^j) \nabla V(X_t^j) + \nabla_{X_t^i} k(X_t^i, X_t^j) \right) dt + \sum_{j=1}^N \sqrt{\frac{2}{N} \mathcal{K}(X_t^1, \dots, X_t^N)} dW_t^j,$$

which is ergodic with respect to the product measure  $\pi^{\otimes N}$ , hence in principle exact for a finite number of particles. While the noise term disappears in the limit as  $N \rightarrow \infty$  [Duncan et al., 2023, Proposition 2], [Nüsken and Renger, 2023], neglecting it clearly leads to an underspread finite-particle posterior approximation. A numerical investigation of those aspects can be found in Section 7.

## 5 On the geometry of Stein transport

In this section we discuss the relationship of Stein transport and SVGD to optimal transport, explaining some of the similarities between (11) and (12) from a conceptual angle. Investigations into the Stein geometry were started by Liu [2017] and have further been developed by Duncan et al. [2023], Nüsken and Renger [2023].

The key object in the Stein geometry is the extended<sup>4</sup> metric  $d_k$  between probability measures  $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ ,

$$d_k^2(\mu, \nu) = \inf_{(\pi_t, v_t)_{t \in [0,1]}} \left\{ \int_0^1 \|v_t\|_{\mathcal{H}_k^d}^2 dt : \quad \partial_t \pi_t + \nabla \cdot (\pi_t v_t) = 0, \quad \pi_0 = \mu, \quad \pi_1 = \nu \right\}. \quad (26)$$

Intuitively speaking, the continuity equation  $\partial_t \pi_t + \nabla \cdot (\pi_t v_t) = 0$  asserts that the probability measure  $\pi_t$  is transported by the vector fields  $v_t$ , in the sense of the ordinary differential equation (ODE)  $dX_t = v_t(X_t) dt$ , with random initial condition  $X_0 \sim \pi_0$ , see Santambrogio [2015, Section 4.1.2]. Therefore,  $d_k$  measures the distance between  $\mu$  and  $\nu$  in terms of the length of the shortest connecting curve (geodesic) of probability measures (hence  $d_k$  may be thought of as a geodesic distance in a Riemannian manifold), when the (infinitesimal) length of curves is measured in terms of the RKHS-norms of their driving vector fields. A key motivation for considering (26) is that replacing  $\|\cdot\|_{\mathcal{H}_k^d}$  by  $\|\cdot\|_{L^2(\pi_t)}$  in (26) recovers the Benamou-Brenier representation of the quadratic Wasserstein distance  $d_{W_2}$  [Benamou and Brenier, 2000]; a detailed comparison between  $d_k$  and  $d_{W_2}$  can be found in Duncan et al. [2023, Appendix A].

According to Liu [2017, Theorem 3.5], SVGD performs gradient flow dynamics of the Kullback-Leibler divergence (KL) with respect to  $d_k$ ,

$$\partial_t \pi_t = -\text{grad}_k \text{KL}(\pi_t | \pi_1), \quad (27)$$

recalling that  $\pi_1$  represents the target posterior in our notation. The gradient operation  $\text{grad}_k$  is induced by the geometry encoded by (26), the main point being that the abstract evolution equation (27) governs the mean-field limit of (12) under this interpretation [Liu, 2017, Duncan et al., 2023]. In order to obtain a similar geometric characterisation of Stein transport (that is, of the coupled system composed of (16) and (22)), we notice that (26) naturally induces a notion of *Stein optimal transport maps* (namely those that arise from the shortest connecting curve):

**Definition 1** (Stein optimal transport maps). *Let  $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$  be two probability measures such that  $d_k(\mu, \nu) < \infty$ . Assume that  $(\pi_t, v_t)_{t \in [0,1]} \in C^1([0,1] \times \mathbb{R}^d; \mathbb{R}) \times C_b^1([0,1] \times \mathbb{R}^d; \mathbb{R}^d)$  is a geodesic, that is, a connecting curve that realises the infimum in (26):*

$$\partial_t \pi_t + \nabla \cdot (\pi_t v_t) = 0, \quad \pi_0 = \mu, \quad \pi_1 = \nu, \quad \text{and} \quad \int_0^1 \|v_t\|_{\mathcal{H}_k^d}^2 dt = d_k^2(\mu, \nu). \quad (28)$$

The corresponding ODE  $\frac{dX_t}{dt} = v_t(X_t)$  induces a time-one flow map<sup>5</sup>  $F : \mathbb{R}^d \rightarrow \mathbb{R}^d$  which we then call a Stein optimal transport map between  $\mu$  and  $\nu$ .

**Remark 10** (Wasserstein geodesics). As alluded to above, replacing  $\|\cdot\|_{\mathcal{H}_k^d}$  by  $\|\cdot\|_{L^2(\pi_t)}$  in (26) recovers the Wasserstein-2 distance via the Benamou-Brenier formula [Benamou and Brenier, 2000]. If Definition 1 is modified accordingly (replacing Stein geodesics by Wasserstein geodesics), then the induced optimal transport maps coincide with the conventional ones for the quadratic cost [Villani, 2003, Section 8.1].

By definition, Stein optimal transport maps in the sense of Definition 1 satisfy  $F_\# \mu = \nu$ , where  $F_\#$  denotes the pushforward of measures. As pointed out by El Moseley and Marzouk [2012], constructing a transport map  $F$

<sup>4</sup>An extended metric satisfies the usual axioms of a metric, but may take the value infinity.

<sup>5</sup>The time-one flow map associated to an ODE maps initial conditions to the corresponding solution at time  $t = 1$ , that is,  $F(x_0) = X_1$  if  $X_t$  solves the ODE  $\frac{dX_t}{dt} = v_t(X_t)$  with initial condition  $X_0 = x_0$ .

connecting prior and posterior (that is,  $F_{\#}\pi_0 = \pi_1$ ) would solve the Bayesian inference problem, as then samples from the prior  $\pi_0$  can be transformed into samples from the posterior  $\pi_1$  (in formulas:  $X_0 \sim \pi_0 \implies F(X_0) \sim \pi_1$ ). However, obtaining geodesics for (26) is computationally demanding; indeed the first order minimality conditions are given by the following (numerically challenging) system of coupled partial differential equations [Duncan et al., 2023, Proposition 18],

$$\partial_t \pi_t + \nabla \cdot (\pi_t \mathcal{T}_{k,\pi_t} \nabla \phi_t) = 0, \quad (29a)$$

$$\partial_t \phi_t + \nabla \phi_t \cdot \mathcal{T}_{k,\pi_t} \nabla \phi_t = 0, \quad (29b)$$

recalling the convolution-type integral operator  $\mathcal{T}_{k,\pi}$  defined in (15).

As we show in the remainder of this section, the problem becomes tractable when  $\mu$  and  $\nu$  are close, and thus the transport is infinitesimal in nature. In this regime, it will turn out that it is sufficient to solve (29a), and that (29b) can be dispensed with. Stein transport approximately solves (29a); we summarise this finding as follows:

**Informal result** (Optimal transport interpretation). *After partitioning the time interval  $[0, 1]$  using the discretisation  $0 = t_0 < t_1 < \dots < t_{N_{\text{steps}}} = 1$ , Stein-transport implements infinitesimally optimal Stein transport maps (or geodesic flow) between  $\pi_{t_i}$  and  $\pi_{t_{i+1}}$ , successively for  $i = 0, \dots, N_{\text{steps}} - 1$ . Here, the marginals  $\pi_{t_i}$  are fixed by the interpolation (1).*

Before making this statement precise in Theorem 6 below, we present an informal derivation of the update equations (11), starting from the geodesic equations (29):

**Alternative derivation of Stein-transport.** For the Stein optimal transport between  $\pi_{t_i}$  and  $\pi_{t_{i+1}}$ , we may assume that  $\phi_t$  is constant on the small time interval  $[t_i, t_{i+1}]$ . We therefore concentrate on (29a) rather than (29b) (which is an update equation for  $\phi_t$ ) and impose the evolution equation for  $\pi_t$  as dictated by the interpolation (1). The time derivative satisfies

$$\partial_t \pi_t = -\pi_t \left( h - \int_{\mathbb{R}^d} h d\pi_t \right), \quad (30)$$

see the proof of Proposition 1 in Appendix A. Equating (29a) and (30), we obtain the Stein-Poisson equation (16), up to the regularising term  $\lambda \phi \pi$ . To arrive at (11), notice that we can write

$$-\nabla \cdot (\pi_t \mathcal{T}_{k,\pi_t} \nabla \phi_t) = \int_{\mathbb{R}^d} \xi^{k,\nabla \log \pi_t}(\cdot, y) \phi(y) \pi(dy),$$

performing integration by parts and using the definition (6) of the KSD-kernel  $\xi^{k,\nabla \log \pi}$ . Formally replacing  $\pi$  by the empirical measure  $\frac{1}{N} \sum_{i=1}^N \delta_{X_t^i}$ , and adding the regularisation term involving  $\lambda$  leads to (11), cf. Remark 5.

To make this informal discussion precise we present Proposition 6 below, which establishes that Stein optimal transport is indeed governed by (29a) when  $\mu$  and  $\nu$  are close. The ‘closeness’ between probability distributions is captured rigorously by considering (short) curves and differentiating at zero (corresponding to linearisation or first-order Taylor expansions).

**Proposition 6** (Infinitesimal Stein transport and the Stein-Poisson equation). *Let  $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  be bounded, with bounded first order derivatives. Fix  $\varepsilon > 0$  and let  $(\pi_t)_{t \in (-\varepsilon, \varepsilon)} \subset \mathcal{P}(\mathbb{R}^d)$  be a curve of probability measures such that  $t \mapsto d_k(\pi_0, \pi_t)$  is Lipschitz-continuous. Assume that the Stein optimal transport maps between  $\pi_0$  and  $\pi_t$  exist and are unique, for all  $t \in (-\varepsilon, \varepsilon)$ , and denote those maps by  $F_t$ . Assume moreover that there exist solutions  $\phi_t \in L^2(\pi_t)$  to the Stein-Poisson equations*

$$\nabla \cdot (\pi_t \mathcal{T}_{k,\pi_t} \nabla \phi_t) = -\partial_t \pi_t, \quad (31)$$

where the time-derivatives  $\partial_t \pi_t$  are assumed to exist in the distributional sense, and that  $(-\varepsilon, \varepsilon) \ni t \mapsto \mathcal{T}_{k,\pi_t} \nabla \phi_t \in \mathcal{H}_k^d$  is continuous.

Then we have

$$\frac{d^+}{dt} F_t(x) := \lim_{t \rightarrow 0^+} \frac{F_t(x) - x}{t} = \mathcal{T}_{k,\pi_0} \nabla \phi_0.$$

*Proof.* See Appendix C. The proof is inspired by the proof of Proposition 8.4.6 in Ambrosio et al. [2008], which is an analogous result for the  $W_p$ -Wasserstein distance.  $\square$

## 6 Other perspectives and related work

### 6.1 Regression and transport

Score and flow matching [Yang et al., 2023, Song et al., 2021, Lipman et al., 2023, Liu et al., 2023] have recently led to breakthroughs in generative modeling, and, based on conditional expectations, both approaches can be framed as regression type approximations of vector fields in dynamical models, not unlike (5). Stochastic optimal control is also related to least-squares formulations via backward stochastic differential equations (BSDEs), see, for example, [Pham, 2009, Chapter 6], Chessari et al. [2023] or [Richter et al., 2023, Section 2].

Interacting particle systems similar to Stein transport have been constructed in recent works [Maurais and Marzouk, 2023, 2024, Wang and Nüsken, 2024], named *kernelised Fisher-Rao (KFR) flow* or *kernel mean embedding (KME) dynamics*. From a regression perspective, those can be understood in terms of the minimisation problem [Wang and Nüsken, 2024, Section 4]

$$v^* \in \underset{v \in (L^2(\pi))^d}{\operatorname{argmin}} \left( \|\mathcal{T}_{k,\pi}(S_\pi v - h_0)\|_{\mathcal{H}_k}^2 + \lambda \|v\|_{(L^2(\pi))^d}^2 \right) \quad (32a)$$

$$= \underset{v \in (L^2(\pi))^d}{\operatorname{argmin}} \left( \text{MMD}_k(-\nabla \cdot (\pi v), -\pi h_0) + \lambda \|v\|_{(L^2(\pi))^d}^2 \right), \quad (32b)$$

where  $\text{MMD}_k(-\nabla \cdot (\pi v), -\pi h_0)$  refers to the maximum mean discrepancy [Smola et al., 2007, Gretton et al., 2012, Muandet et al., 2017] between the signed measures  $-\nabla \cdot (\pi v)$ , induced by the flow of the ODE  $dX_t = v(X_t) dt$ , and  $-\pi h_0 = \partial_t \pi_t$  given by the interpolation (1). The construction in (32b) aims to match flows based on their kernel mean embeddings, while (32a) is instructive in light of its comparison to (14): the Stein equation (3) is embedded into  $\mathcal{H}_k$  via  $\mathcal{T}_{k,\pi}$  and the roles of  $\|\cdot\|_{\mathcal{H}_k}$  and  $\|\cdot\|_{L^2\pi}$  have been reversed. In practical terms, KFR/KME-flow replaces the vector fields  $\nabla_{X^i} k(\cdot, X^i) + k(\cdot, X^i) \nabla \log \pi_t(X^i)$  by  $\nabla_{X^i} k(\cdot, X^i)$ : it does not leverage or require the scores  $\nabla \log \pi_t$ , and is therefore applicable in scenarios where the prior  $\pi_0$  is only available through a sample-based approximation (as is typical in data assimilation [Law et al., 2015], for instance). On the other hand, in situations where the score can be evaluated, it is expected that methods incorporating this information into the inference procedure perform better in terms of accuracy and scalability.

SVGD has been connected to problems of regression type through the formulation

$$v^* \in \underset{v \in \mathcal{H}_k^d}{\operatorname{argmin}} \left( \int_{\mathbb{R}^d} \left( v - \nabla \log \left( \frac{d\pi}{d\pi_1} \right) \right)^2 d\pi + \lambda \|v\|_{\mathcal{H}_k^d}^2 \right), \quad (33)$$

where the optimal velocity field  $v^* = (\mathcal{T}_{k,\pi} + \lambda I)^{-1} \mathcal{T}_{k,\pi} \left( \nabla \log \frac{d\pi}{d\pi_1} \right)$  is a close relative of the SVGD velocity field  $v_{\text{SVGD}} = \mathcal{T}_{k,\pi} \left( \nabla \log \frac{d\pi}{d\pi_1} \right)$  and recovers the Wasserstein-2 gradient  $\nabla \log \frac{d\pi}{d\pi_1}$  in the limit as  $\lambda \rightarrow 0$  [Maoutsa et al., 2020, He et al., 2024, Zhu and Mielke, 2024]. From an algorithmic perspective, interacting particle systems based on (33) have the same fixed points as standard SVGD, and are therefore unlikely to overcome the finite-particle issues associated with SVGD (see Section 7.2.2 and Remark 9).

### 6.2 Gradient flows

In Section 5, we have shown that Stein transport performs infinitesimally optimal Stein transport maps between  $\pi_t$  and  $\pi_{t+\Delta t}$ : this is a statement at the particle level, or, in other words, about couplings between  $\pi_t$  and  $\pi_{t+\Delta t}$ . Moreover, the construction does not depend on the specific form of the interpolation (1), as explained in Remark 1. However, at the level of densities, the interpolation (1) can be understood from a geometrical perspective as follows: Define the time-reparameterised curve  $\rho_t := \pi_{-\log(1-t)}$  as in Domingo-Enrich and Pooladian [2023], Chopin et al. [2023], satisfying  $\rho_0 = \pi_0$  and  $\lim_{t \rightarrow \infty} \rho_t = \pi_1$ . A direct calculation shows that

$$\partial_t \rho_t = -\rho_t \left( \log \left( \frac{\rho_t}{\pi_1} \right) - \int_{\mathbb{R}^d} \log \left( \frac{\rho_t}{\pi_1} \right) d\rho_t \right), \quad (34)$$

and this equation can be understood as a KL-gradient flow in the Fisher-Rao geometry [Lu et al., 2019b, 2023, Zhu and Mielke, 2024, Chen et al., 2023a,b], or as a mirror gradient flow [Domingo-Enrich and Pooladian, 2023, Chopin et al., 2023]. However, our impression is that the more salient point is that (34) can be understood as a natural or Newton-type gradient flow [Amari, 2016, Chapter 12],

$$\partial_t \rho_t = (\text{Hess KL}(\rho_t | \pi_1))^{-1} \nabla \text{KL}(\rho_t | \pi_1), \quad (35)$$

where the Hessian and the gradient are interpreted in the ‘Euclidean’ way, that is, in the geometry generated by ‘vertical’ geodesics of the form  $\rho_t = (1-t)\rho_0 + t\rho_1$ , for  $t \in [0, 1]$ . The formulation (35) is essentially a corollary of the mirror flow perspective (see, for example, [Kerimkulov et al., 2023, Appendix B] or [Raskutti and Mukherjee, 2015, Theorem 1]), but for the convenience of the reader, we provide a self-contained explanation in Appendix D.

Since the Hessian in (35) acts as a preconditioner, it is reasonable to expect that methods based on the interpolation (1) do not suffer significantly from ill-conditioned targets (featuring, for example, elongated or distorted modes with an intricate geometry), and indeed convergence in (35) is independent of the log-Sobolev constant of the target  $\pi_1$  [Lu et al., 2019b, 2023, Domingo-Enrich and Pooladian, 2023, Chen et al., 2023a]. To corroborate this intuition, we numerically compare Stein transport to SVGD in Section 7.2.1 on a target that exhibits the above mentioned characteristics. Finally, it is interesting to remark that Newton-type gradient flows of the form (35) have been constructed with  $\text{Hess KL}(\rho_t|\pi_1)$  and  $\nabla \text{KL}(\rho_t|\pi_1)$  derived instead from the Wasserstein geometry [Wang and Li, 2020] or the Stein geometry [Detommaso et al., 2018], but those schemes, in contrast to Stein transport, require access to  $\text{Hess log } \pi_1$ , the Hessian of the log-target.

## 7 Implementation and numerical experiments

### 7.1 Adjusted Stein transport

In this section, we detail a modification to the implementation of Algorithm 1 that greatly stabilises the method and substantially improves numerical performance. The accuracy of a single Stein transport step (one Euler step on the continuous-time system (11)) depends crucially on the accuracy of the approximation  $\pi_{t_i} \approx \frac{1}{N} \sum_{j=1}^N \delta_{X_{t_i}^j}$ . In the context of Proposition 5, this translates into properties of the projection  $P_X$  that should be as close to the identity on  $\mathcal{H}_k^d$  as possible. Additionally, if two particles are very close to each other,  $X_{t_i}^j \approx X_{t_i}^l$  for some  $j \neq l$ , then inversion of the linear system in (11) becomes highly unstable, as the Gram matrix  $\xi^{k, \nabla \log \pi_{t_i}}$  associated to the KSD-kernel  $\xi^{k, \nabla \log \pi_{t_i}}$  becomes nearly degenerate. Errors of this type tend to accumulate as the algorithm progresses. To mitigate these problems, we find it beneficial to apply a few SVGD steps targeting  $\pi_{t_i}$  before carrying out the Stein transport  $\pi_{t_i} \mapsto \pi_{t_{i+1}}$ . This SVGD-adjustment improves the approximation  $\pi_{t_i} \approx \frac{1}{N} \sum_{j=1}^N \delta_{X_{t_i}^j}$ , and the repulsive term  $\nabla_{X_t^j} k(X_t^i, X_t^j)$  in (12) separates the particles, thereby improving the conditioning of the linear system in (11). We would like to stress that in the suggested algorithm, SVGD is not used to target the actual posterior  $\pi_1$ ; the method is rather very much akin to the use of MCMC samplers within a sequential Monte Carlo framework [Chopin and Papaspiliopoulos, 2020, Chapter 17]. For clarity, we summarise SVGD-adjusted Stein transport in Algorithm 2. Notice that the incorporation of SVGD steps is fairly adhoc, and a more systematic study of adjustment schemes (possibly relating to the optimal design of experiments [Huan et al., 2024]) could be an interesting direction for future work.

---

**Algorithm 2** Adjusted Stein transport

---

**Require:** Prior samples  $X_0^1, \dots, X_0^N \in \mathbb{R}^d$ , prior score function  $\nabla \log \pi_0 : \mathbb{R}^d \rightarrow \mathbb{R}^d$ , negative log-likelihood  $h : \mathbb{R}^d \rightarrow \mathbb{R}$ , with gradient  $\nabla h : \mathbb{R}^d \rightarrow \mathbb{R}^d$ , positive definite kernel  $k$ , regularisation parameter  $\lambda > 0$ , time discretisation  $0 = t_0 < t_1 < \dots < t_{N_{\text{steps}}} = 1$  with time step  $\Delta t$ , number  $N_{\text{adjust}}$  of interspersed SVGD steps per transport step, SVGD time step  $\Delta t_{\text{adjust}}$

```

for  $n = 0, \dots, N_{\text{steps}} - 1$  do
    for  $n = 0, \dots, N_{\text{SVGD}} - 1$  do
         $X_n^i \leftarrow X_n^i + \frac{\Delta t_{\text{adjust}}}{N} \sum_{j=1}^N \left( k(X_n^i, X_n^j) \nabla \log \pi_{t_n}(X_n^j) + \nabla_{X_n^j} k(X_n^i, X_n^j) \right)$ .
    end for
    Compute the scores  $P_n^j = -t_n \nabla h(X_n^j) + \nabla \log \pi_0(X_n^j)$ .
    Compute  $\xi_{ij} = P_n^i \cdot \nabla_{X_n^j} k(X_n^i, X_n^j) + P_n^j \cdot \nabla_{X_n^i} k(X_n^i, X_n^j)$ 
         $+ \nabla_{X_n^i} \cdot \nabla_{X_n^j} k(X_n^i, X_n^j) + P_n^i \cdot k(X_n^i, X_n^j) P_n^j$ .
    Compute  $\mathbf{h}_i = h(X_n^i) - \frac{1}{N} \sum_{i=1}^N h(X_n^i)$ .
    Solve  $(\xi + \lambda I_{N \times N})\phi = \mathbf{h}$  for  $\phi \in \mathbb{R}^N$ .
     $X_{n+1}^i \leftarrow X_n^i + \frac{\Delta t}{N} \sum_{j=1}^N \phi^j \left( k(X_n^i, X_n^j) P_n^j(X_n^j) + \nabla_{X_n^j} k(X_n^i, X_n^j) \right)$ .
end for
return posterior samples  $X_{N_{\text{steps}}}^1, \dots, X_{N_{\text{steps}}}^N$ .

```

SVGD-  
adjustment  
targeting  $\pi_{t_n}$ .
Stein transport  
step from  
Algorithm 1.

## 7.2 Numerical experiments

In the following, we compare Stein transport (Algorithm 1), its SVGD-adjusted variant (Algorithm 2) and SVGD in a number of test cases. For all methods, we use the square-exponential kernel

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right), \quad (36)$$

where the bandwidth  $\sigma^2$  is chosen adaptively according to the median heuristic [Liu and Wang, 2016],  $\sigma^2 = \text{med}^2/(2 \log N)$ , with med being the median of the pairwise Euclidean distance between the current particle positions. SVGD is implemented using the Adagrad optimiser as suggested by Liu and Wang [2016].

### 7.2.1 Implicit preconditioning: the Joker distribution

We follow Detommaso et al. [2018, Section 5.1] and consider sampling from a two-dimensional Bayesian posterior, derived from the forward operator

$$\mathcal{F}(x) = \log((1 - x_1)^2 + 100(x_2 - x_1^2)^2), \quad (x_1, x_2) = x,$$

a scalar logarithmic Rosenbrock function. We impose the prior  $x \sim \mathcal{N}(0, I_{2 \times 2})$ , and collect a single observation  $y_{\text{obs}} = \mathcal{F}(x_{\text{true}}) + \xi$ , with  $\xi \sim \mathcal{N}(0, \sigma^2)$  and  $x_{\text{true}}$  drawn from the prior, inducing the (unnormalised) likelihood  $\exp(-\frac{1}{2\sigma^2} \|\mathcal{F}(x) - y_{\text{obs}}\|^2)$ . In other words, the posterior is given by

$$\pi(x) \propto \exp\left(-\frac{1}{2\sigma^2} \|\mathcal{F}(x) - y_{\text{obs}}\|^2 - \frac{1}{2} \|x\|^2\right),$$

and we plot its density (for  $\sigma = 0.3$ ) in Figure 1a. We observe that the contours of the target are fairly sharp and narrow, and indeed Detommaso et al. [2018] use this example to exhibit the benefits of appropriately preconditioned versions of SVGD (cf. the discussion in Section 6.2). We run Stein transport, SVGD, and adjusted Stein transport, initialising  $N = 500$  particles from the prior. For Stein transport and the adjusted variant, we partition the time interval  $[0, 1]$  into  $N_{\text{steps}} = 50$  equidistant steps and use the regularisation parameter  $\lambda = 10^{-2}$  (see Algorithm 1). Adjusted Stein transport (Algorithm 2) is implemented with one SVGD step per transport step, with step size  $\Delta_{\text{adjust}} = 0.02$  (this is the same step size as for the transport step), and the same regularisation  $\lambda = 10^{-2}$ . SVGD is run for 250 steps with step size  $\Delta t = 0.01$ , and using Adagrad updates as in Liu and Wang [2016]. All methods rely on the square-exponential kernel (36), with dynamically selected bandwidth according to the median heuristic [Liu and Wang, 2016].

Figure 1 shows the posterior approximations obtained by the different methods. In Figure 2, we show the kernelised Stein discrepancy (KSD) towards the target  $\pi$ , as a function of the iteration count (or rather, as a function of the number of evaluations of  $\nabla h$ ). As suggested by Gorham and Mackey [2017], KSD is implemented using the inverse multiquadric kernel  $k_{\text{IMQ}}(x, y) = (1 + \|x - y\|^2)^{-1/2}$ . Notice that adjusted Stein transport takes twice as many gradient evaluations compared to the unadjusted counterpart, due to the interspersed SVGD steps. Figure 1b shows that (unadjusted) Stein transport is only able to provide a rather inaccurate approximation of the posterior. As discussed in Section 7.1, this observation can be attributed to instabilities and accumulation of errors. Indeed, Figure 2 shows that unadjusted Stein transport fails to improve the KSD-score after roughly half of the steps, an indication that the particle representation of the intermediate distributions has become unreliable (or the Gram matrix  $\xi$  has degenerated) and thus Stein transport has become ineffective. Comparing Figures 1c and 1d, we observe that adjusted Stein transport is able to fit the sharp contour lines of the posterior more accurately than SVGD, resulting in a lower final KSD-score (see Figure 2). We attribute this finding to the fact that (adjusted) Stein transport follows a Newton-type gradient flow (see Section 6.2), and thus our results are in line with those of Detommaso et al. [2018] for the Stein variational Newton method. Notice that, in contrast to Stein variational Newton, adjusted Stein transport does not require evaluations of the Hessian of the log-target.

### 7.2.2 (Non-)Collapse in high dimensions

In this section, we perform a numerical investigation of the variance under- and overestimation phenomenon of SVGD and Stein transport, respectively (see Remark 9 and the discussion around Proposition 5). We also refer to Ba et al. [2021], Gong et al. [2021], Liu et al. [2022], Zhuo et al. [2018] for background.

**Multivariate Gaussians.** In a first experiment, we consider sampling from a multivariate Gaussian target, a standard example that illustrates the variance collapse of SVGD in high dimensions [Ba et al., 2021, Gong et al., 2021, Liu et al., 2022, Zhuo et al., 2018]. We initialise the particles from the Gaussian prior  $\mathcal{N}(\mathbf{1}, I_{d \times d})$ , where  $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^d$  denotes the  $d$ -dimensional ‘all ones’ vector. With the negative log-likelihood  $h(x) = \frac{1}{2} \|x + \mathbf{1}\|^2$  (that is, we impose a Gaussian observation model with unit covariance and observation  $y_{\text{obs}} = -\mathbf{1}$ ), the target posterior is  $\mathcal{N}(0, \frac{1}{2} I_{d \times d})$ . We

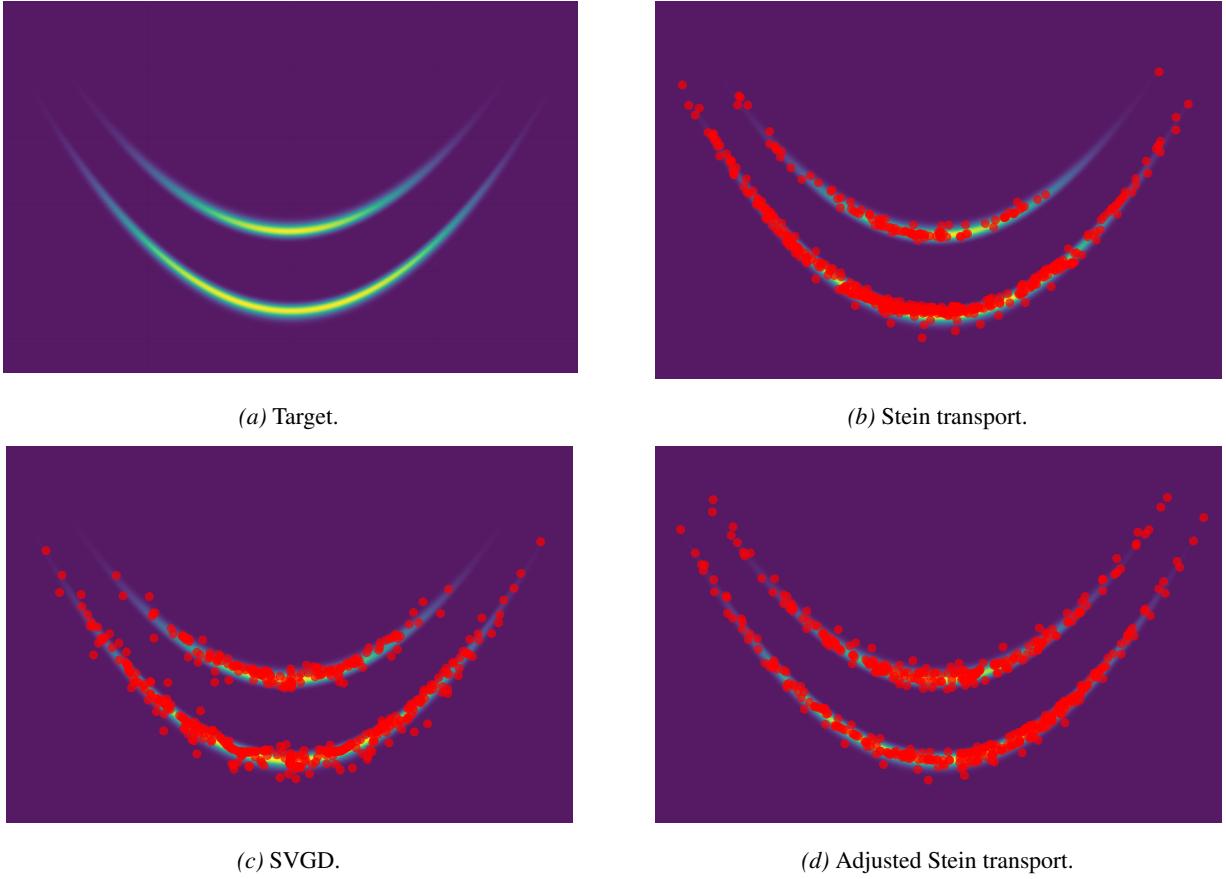


Figure 1. Posterior approximations for the Joker distribution.

initialise 200 particles from the prior, and run SVGD (using the Adagrad optimiser with step size  $\Delta t = 0.1$  for 200 steps) as well as Stein transport (for 100 steps and regularisation  $\lambda = 10^{-2}$ ). The adjusted variant is run with 5, 10, 20 or 100 interspersed SVGD steps, each of which with step size  $\Delta t_{\text{adjust}} = 0.1$  and using Adagrad.<sup>6</sup>

In Figure 3, we plot  $\frac{1}{d} \text{Tr} \widehat{\text{Cov}}$ , the trace of the estimated covariance matrix, rescaled by  $\frac{1}{d}$ , as a function of  $d$ . The black line indicates the true value,  $\frac{1}{d} \text{Tr} \text{Cov} \mathcal{N}(0, \frac{1}{2} I_{d \times d}) = \frac{1}{2}$ . In line with prior works and Remark 9, SVGD severely underestimates the posterior variance in high-dimensional scenarios. In contrast (and in line with Proposition 5), unadjusted Stein transport severely overestimates the posterior variance. Adjusted Stein transport provides a fairly accurate value for the posterior variance, especially if the number of interspersed SVGD-adjustment steps is large enough ( $\approx 20$ ). We have included the plot for 100 adjustment steps in order to show that the performance saturates; finding an appropriate balance between adjustment and transport moves does not seem to be an issue.

**Low-rank Gaussian mixture.** To further showcase the ability of adjusted Stein transport to avoid posterior collapse in high dimensional settings, we consider the task of sampling from a Gaussian mixture with low-rank structure, following Liu et al. [2022]. More specifically, the target is given by  $\pi(x) = \frac{1}{4} \sum_{i=1}^4 \mathcal{N}(x; \mu_i, I_{d \times d})$ , where the means are defined as

$$\mu_j = \left( \sqrt{5} \cos(2j\pi/4 + \pi/4), \sqrt{5} \cos(2j\pi/4 + \pi/4), 0, \dots, 0 \right)^T \in \mathbb{R}^d,$$

for  $j = 1, \dots, 4$ . The low-rank structure manifests itself in the first two coordinates: the means are placed on a circle, in an equidistant way (see Figure 4). The marginal in the remaining  $d - 2$  coordinates is standard Gaussian, so that when initialising the particles from the prior  $\mathcal{N}(0, I_{d \times d})$ , only the first two coordinates need to be shifted.

We set  $d = 50$ , and implement SVGD and adjusted Stein transport with  $N = 200$  particles. To be specific, SVGD is run for 150 Adagrad steps (with step size  $\Delta t = 0.01$ ) and adjusted Stein transport is run for 100 steps (that is, the step size

<sup>6</sup>Note that using Adagrad here instead of plain SVGD is a slight departure from Algorithm 2.

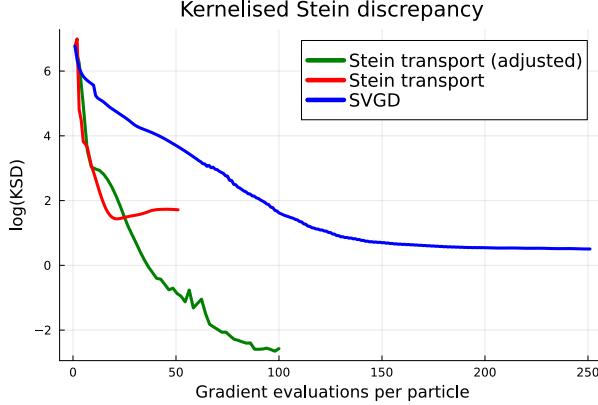


Figure 2. KSD evolution for the Joker distribution from Section 7.2.1.

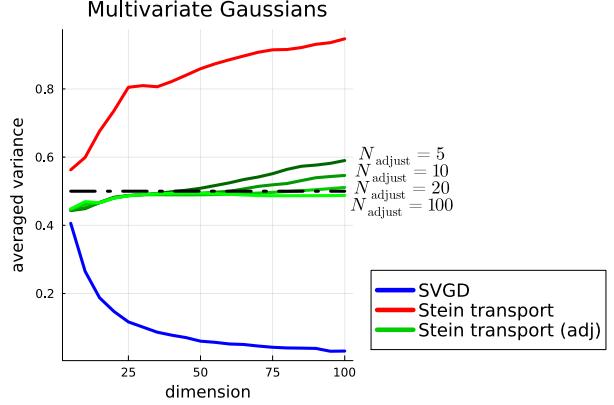


Figure 3. Averaged variance  $\frac{1}{d} \text{Tr} \widehat{\text{Cov}}$  of the approximate posterior, as a function of the dimension. The black line indicates the true value  $\frac{1}{2}$ .

is  $\Delta t = 0.01$ ), interspersed with  $N_{\text{adjust}} = 20$  SVGD Adagrad adjustment steps of size  $\Delta t_{\text{adjust}} = 0.01$ . As we can see in Figure 4a and already observed by Liu et al. [2022], SVGD severely underestimates the spread of the Gaussian mixture components. Adjusted Stein transport, on the other hand, reaches a satisfactory posterior approximation (note that an ensemble size of 200 particles is relatively small in 50 dimensions). In this example, the adjustment steps proved to be crucial; we found it difficult to obtain satisfactory results with unadjusted Stein transport due to instabilities and accumulation of errors.

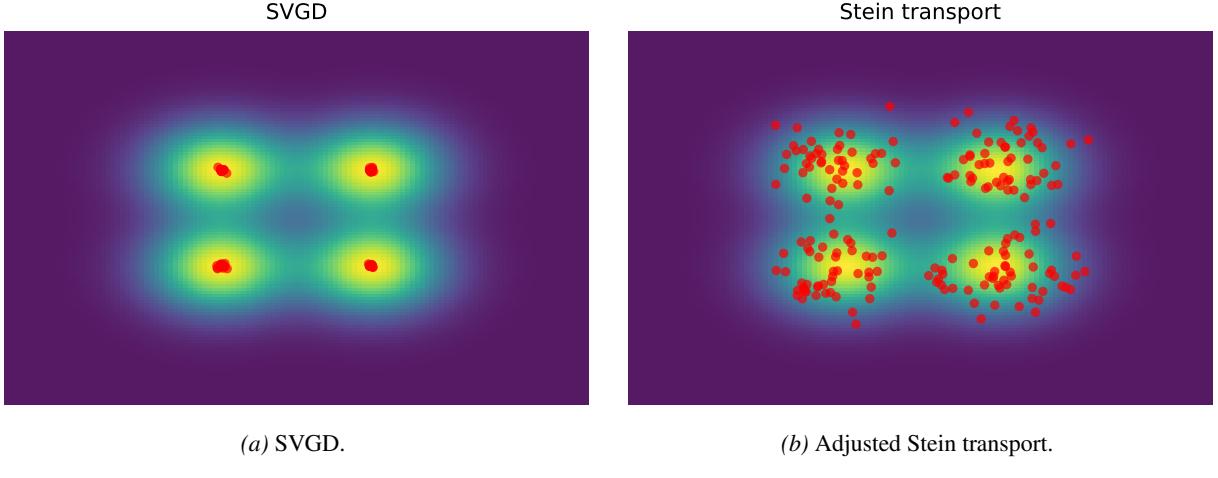


Figure 4. Gaussian mixture with low-rank structure ( $d = 50$ , ensemble size 200 particles). We show the marginals in the first two coordinates of the approximations obtained by SVGD and adjusted Stein transport.

### 7.2.3 Convergence in unit time: Bayesian logistic regression

In our last experiment, we compare SVGD and (adjusted) Stein transport in the context of a Bayesian logistic regression task. We use the 60-dimensional Splice dataset [Rätsch et al., 2001], assume a standard Gaussian prior, and use  $N = 500$  particles. SVGD is run using the Adagrad optimiser with standard parameters and time step  $\Delta t = 0.01$  (significantly increasing the step size leads to instabilities). Adjusted Stein transport is run using 50 steps (that is, with time step  $\Delta t = 0.02$ ) and one SVGD-adjustment step (with time step  $\Delta t = 0.01$ ) per transport step. Figure 5 shows the time evolution of the KSD (using the inverse multiquadric kernel  $k_{\text{IMQ}}(x, y) = (1 + \|x - y\|_2^2)^{-1/2}$  as in Section 7.2.1) and the test accuracy along the dynamics of the particle systems. We observe that adjusted Stein transport reaches low KSD-scores and high test accuracies with significantly less gradient evaluations per particle. We take this observation as an indication that the construction principle behind Stein transport (converging to the posterior at time  $t = 1$  vs  $t \rightarrow \infty$  for SVGD) can indeed significantly reduce the computational cost.

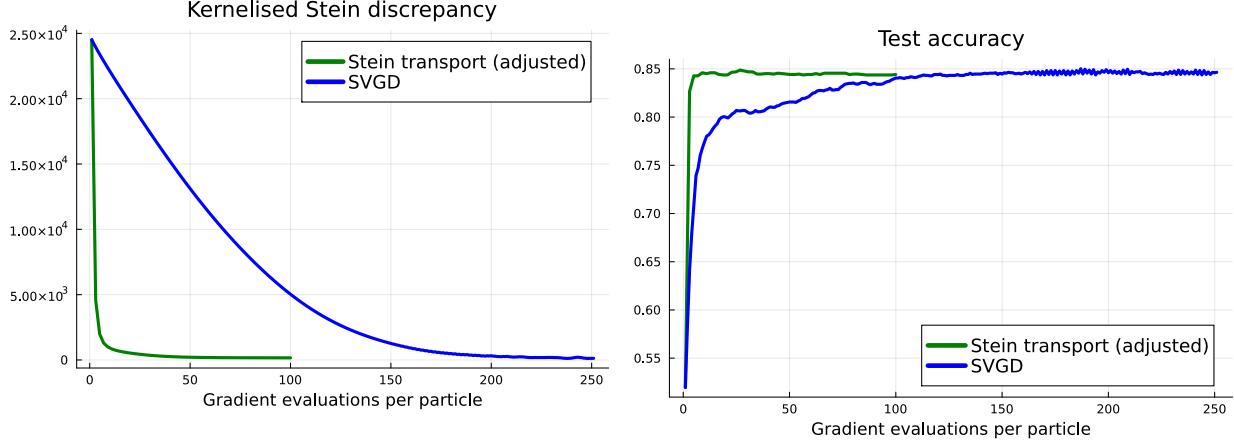


Figure 5. Bayesian logistic regression for the Splice data set ( $d = 60$ ): KSD and test accuracy along the time evolution of the particle system.

## 8 Conclusion, limitations and outlook

We have developed Stein transport based on a hybrid framework that combines regression with dynamics, resulting in a scheme reminiscent of SVGD and sharing its geometrical foundations. The adjusted variant in particular demonstrates promising numerical results that align closely with the underlying theory. Firstly, Stein transport is designed to reach its posterior approximation at time  $t = 1$ ; in our experiments, it indeed achieved similar or superior accuracy compared to SVGD while requiring significantly less computational effort (see in particular Section 7.2.3). Secondly, since Stein transport is constructed to follow a preconditioned gradient flow (see equation (35)), it produces a more accurate posterior approximation in scenarios where the target distribution exhibits complex geometry. Thirdly, unlike SVGD, which often suffers from particle collapse in high-dimensional settings, Stein transport maintains robustness against such issues, arguably because it is constructed from a regression perspective (see Section 7.2.2). This construction ensures that the finite-particle vector field remains close to a projection derived from a mean-field limit (see Proposition 5). In future work, it would be promising to further explore the theoretical insights into regularisation and finite-particle effects (see Section 4.2), and to deepen the connections between statistical estimation and transport methods more broadly. Specifically, it would be valuable to systematically develop weighted schemes and incorporate principled adjustment mechanisms.

**Acknowledgements.** This work was partly supported by Deutsche Forschungsgemeinschaft (DFG) through the grant CRC 1114 ‘Scaling Cascades in Complex Systems’ (project A02, project number 235221301, second funding phase). Many thanks to Deniz Akyildiz for very valuable comments on a preliminary version of this manuscript!

## References

- S.-i. Amari. *Information geometry and its applications*, volume 194. Springer, 2016.
- L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.
- A. Anastasiou, A. Barp, F.-X. Briol, B. Ebner, R. E. Gaunt, F. Ghaderinezhad, J. Gorham, A. Gretton, C. Ley, Q. Liu, et al. Stein’s method meets computational statistics: A review of some recent developments. *Statistical Science*, 38(1):120–139, 2023.
- M. Arbel, A. Korba, A. Salim, and A. Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- N. Ay, J. Jost, H. Ván Lê, and L. Schwachhöfer. *Information geometry*, volume 64. Springer, 2017.
- J. Ba, M. A. Erdogdu, M. Ghassemi, S. Sun, T. Suzuki, D. Wu, and T. Zhang. Understanding the variance collapse of SVGD in high dimensions. In *International Conference on Learning Representations (ICML)*, 2021.
- J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

- G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.
- D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.
- S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.
- Y. Chen, B. Hosseini, H. Owhadi, and A. M. Stuart. Solving and learning nonlinear PDEs with Gaussian processes. *Journal of Computational Physics*, 447:110668, 2021.
- Y. Chen, D. Z. Huang, J. Huang, S. Reich, and A. M. Stuart. Gradient flows for sampling: mean-field models, Gaussian approximations and affine invariance. *arXiv preprint arXiv:2302.11024*, 2023a.
- Y. Chen, D. Z. Huang, J. Huang, S. Reich, and A. M. Stuart. Sampling via gradient flows in the space of probability measures. *arXiv preprint arXiv:2310.03597*, 2023b.
- J. Chessari, R. Kawai, Y. Shinozaki, and T. Yamada. Numerical methods for backward stochastic differential equations: A survey. *Probability Surveys*, 20:486–567, 2023.
- S. Chewi, T. Le Gouic, C. Lu, T. Maunu, and P. Rigollet. SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- S. Chewi, J. Niles-Weed, and P. Rigollet. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 2024.
- L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018.
- N. Chopin and O. Papaspiliopoulos. *An introduction to sequential Monte Carlo*. Springer, 2020.
- N. Chopin, F. R. Crucinio, and A. Korba. A connection between tempering and entropic mirror descent. *arXiv preprint arXiv:2310.11914*, 2023.
- M. Coghi, T. Nilssen, N. Nüsken, and S. Reich. Rough McKean–Vlasov dynamics for robust ensemble Kalman filtering. *The Annals of Applied Probability*, 33(6B):5693–5752, 2023.
- F. Daum and J. Huang. Particle flow for nonlinear filters with log-homotopy. In *Signal and Data Processing of Small Targets 2008*, volume 6969, page 696918. International Society for Optics and Photonics, 2008.
- F. Daum and J. Huang. Nonlinear filters with particle flow induced by log-homotopy. In *Signal Processing, Sensor Fusion, and Target Recognition XVIII*, volume 7336, page 733603. International Society for Optics and Photonics, 2009.
- F. Daum, J. Huang, and A. Noushin. Exact particle flow for nonlinear filters. In *Signal processing, sensor fusion, and target recognition XIX*, volume 7697, page 769704. International society for optics and photonics, 2010.
- E. De Vito, A. Caponnetto, and L. Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 5:59–85, 2005.
- P. Del Moral, A. Doucet, and A. Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):411–436, 2006.
- G. Detommaso, T. Cui, Y. Marzouk, A. Spantini, and R. Scheichl. A Stein variational Newton method. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- C. Domingo-Enrich and A. Pooladian. An explicit expansion of the Kullback–Leibler divergence along its Fisher–Rao gradient flow. *Trans. Mach. Learn. Res.*, 2023, 2023.
- O. R. Dunbar, A. B. Duncan, A. M. Stuart, and M.-T. Wolfram. Ensemble inference methods for models with noisy and expensive likelihoods. *SIAM Journal on Applied Dynamical Systems*, 21(2):1539–1572, 2022.
- A. Duncan, N. Nüsken, and L. Szpruch. On the geometry of Stein variational gradient descent. *Journal of Machine Learning Research*, 24:1–39, 2023.
- T. A. El Moselhy and Y. M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.
- H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.
- G. E. Fasshauer. *Meshfree approximation methods with MATLAB*, volume 6. World Scientific, 2007.
- M. Fisher, T. Nolan, M. Graham, D. Prangle, and C. Oates. Measure transport with kernel Stein discrepancy. In *International Conference on Artificial Intelligence and Statistics*, pages 1054–1062. PMLR, 2021.
- V. Gallego and D. Insua. Stochastic gradient MCMC with repulsive forces. *arXiv:1812.00071*, 2018.

- A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.
- E. Gladin, P. Dvurechensky, A. Mielke, and J.-J. Zhu. Interaction-force transport gradient flows. *arXiv:2405.17075*, 2024.
- W. Gong, Y. Li, and J. M. Hernández-Lobato. Sliced kernelized Stein discrepancy. In *9th International Conference on Learning Representations (ICLR)*, 2021.
- J. Gorham and L. Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning (ICML)*, pages 1292–1301. PMLR, 2017.
- A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.
- R. Hable and A. Christmann. On qualitative robustness of support vector machines. *Journal of Multivariate Analysis*, 102(6):993–1007, 2011.
- Y. He, K. Balasubramanian, B. K. Sriperumbudur, and J. Lu. Regularized Stein variational gradient flow. *Foundations of Computational Mathematics*, pages 1–59, 2024.
- J. Heng, A. Doucet, and Y. Pokern. Gibbs flow for approximate transport with applications to Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(1):156–187, 2021.
- X. Huan, J. Jagalur, and Y. Marzouk. Optimal experimental design: Formulations and computations. *arXiv:2407.16212*, 2024.
- M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.
- J. Jost and J. Jost. *Riemannian geometry and geometric analysis*, volume 42005. Springer, 2008.
- M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv:1807.02582*, 2018.
- B. Kerimkulov, J.-M. Leahy, D. Siska, L. Szpruch, and Y. Zhang. A Fisher-Rao gradient flow for entropy-regularised Markov decision processes in Polish spaces. *arXiv:2310.02951*, 2023.
- A. Kirsch. *An introduction to the mathematical theory of inverse problems*, volume 120. Springer Nature, 2021.
- T. Komorowski, C. Landim, and S. Olla. *Fluctuations in Markov processes: time symmetry and martingale approximation*, volume 345. Springer Science & Business Media, 2012.
- A. Korba, A. Salim, M. Arbel, G. Luise, and A. Gretton. A non-asymptotic analysis for Stein variational gradient descent. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- A. Korba, P.-C. Aubin-Frankowski, S. Majewski, and P. Ablin. Kernel Stein discrepancy descent. *International Conference on Machine Learning (ICML)*, 2021.
- R. S. Laugesen, P. G. Mehta, S. P. Meyn, and M. Raginsky. Poisson’s equation in nonlinear filtering. *SIAM Journal on Control and Optimization*, 53(1):501–525, 2015.
- K. Law, A. Stuart, and K. Zygalakis. Data assimilation. *Cham, Switzerland: Springer*, 214:52, 2015.
- J. M. Lee and J. M. Lee. *Smooth manifolds*. Springer, 2012.
- T. Lelièvre and G. Stoltz. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numerica*, 25:681–880, 2016.
- M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.
- Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR*, 2023.
- C. Liu, J. Zhuo, P. Cheng, R. Zhang, and J. Zhu. Understanding and accelerating particle-based variational inference. In *International Conference on Machine Learning*, pages 4082–4092. PMLR, 2019.
- Q. Liu. Stein variational gradient descent as gradient flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International conference on machine learning (ICML)*, pages 276–284. PMLR, 2016.

- X. Liu, H. Zhu, J. Ton, G. Wynne, and A. B. Duncan. Grassmann Stein variational gradient descent. In *International Conference on Artificial Intelligence and Statistics, AISTATS*. PMLR, 2022.
- X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.
- J. Lu, Y. Lu, and J. Nolen. Scaling limit of the Stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671, 2019a.
- Y. Lu, J. Lu, and J. Nolen. Accelerating Langevin sampling with birth-death. *arXiv:1905.09863*, 2019b.
- Y. Lu, D. Slepčev, and L. Wang. Birth–death dynamics for sampling: global convergence, approximations and their asymptotics. *Nonlinearity*, 36(11):5731, 2023.
- D. Maoutsa, S. Reich, and M. Opper. Interacting particle solutions of Fokker–Planck equations through gradient–log-density estimation. *Entropy*, 22(8):802, 2020.
- A. Maurais and Y. Marzouk. Adaptive algorithms for continuous-time transport: Homotopy-driven sampling and a new interacting particle system. In *NeurIPS 2023 Workshop Optimal Transport and Machine Learning*, 2023.
- A. Maurais and Y. Marzouk. Sampling in unit time with kernel Fisher-Rao flow. *arXiv:2401.03892*, 2024.
- A. Mielke. An introduction to the analysis of gradients systems. *arXiv preprint arXiv:2306.05026*, 2023.
- K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.
- N. Nüsken and D. Renger. Stein variational gradient descent: Many-particle and long-time asymptotics. *Foundations of Data Science*, 5(3):286–320, 2023.
- C. J. Oates, T. Papamarkou, and M. Girolami. The controlled thermodynamic integral for Bayesian model evidence evaluation. *Journal of the American Statistical Association*, 111(514):634–645, 2016.
- C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):695–718, 2017.
- É. Pardoux and Y. Veretennikov. On the Poisson equation and diffusion approximation. I. *The Annals of Probability*, 29(3):1061–1085, 2001.
- S. Pathiraja, S. Reich, and W. Stannat. McKean–Vlasov SDEs in nonlinear filtering. *SIAM Journal on Control and Optimization*, 59(6):4188–4215, 2021.
- H. Pham. *Continuous-time stochastic control and optimization with financial applications*, volume 61. Springer Science & Business Media, 2009.
- A. Radhakrishnan and S. Meyn. Gain function tracking in the feedback particle filter. In *2019 American Control Conference (ACC)*, pages 5352–5359. IEEE, 2019.
- A. Rahimi, B. Recht, et al. Random features for large-scale kernel machines. In *NIPS*, volume 3, page 5. Citeseer, 2007.
- G. Raskutti and S. Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.
- G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for adaboost. *Machine learning*, 42:287–320, 2001.
- S. Reich. A dynamical systems framework for intermittent data assimilation. *BIT Numerical Mathematics*, 51(1):235–249, 2011.
- S. Reich. A Gaussian-mixture ensemble transform filter. *Quarterly Journal of the Royal Meteorological Society*, 138(662):222–233, 2012.
- S. Reich. Data assimilation: A dynamic homotopy-based coupling approach. In *Stochastic Transport in Upper Ocean Dynamics Annual Workshop*, pages 261–280. Springer Nature Switzerland Cham, 2022.
- L. Richter, L. Sallandt, and N. Nüsken. From continuous-time formulations to discretization schemes: tensor trains and robust regression for BSDEs and parabolic PDEs. *arXiv:2307.15496*, 2023.
- C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.
- A. Rudi, L. Carratino, and L. Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- F. Santambrogio. Optimal transport for applied mathematicians. *Birkhäuser, NY*, 55(58-63):94, 2015.
- S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

- A. Smith. *Sequential Monte Carlo methods in practice*. Springer Science & Business Media, 2013.
- A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *International conference on algorithmic learning theory*, pages 13–31. Springer, 2007.
- A. J. Smola and B. Schölkopf. *Learning with kernels*, volume 4. Citeseer, 1998.
- Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR*, 2021.
- B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.
- I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.
- S. Syed, V. Romaniello, T. Campbell, and A. Bouchard-Côté. Parallel tempering on optimized paths. In *International Conference on Machine Learning (ICML)*, pages 10033–10042. PMLR, 2021.
- A. Taghvaei and P. G. Mehta. Gain function approximation in the feedback particle filter. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 5446–5452. IEEE, 2016.
- A. Taghvaei, P. G. Mehta, and S. P. Meyn. Diffusion map-based algorithm for gain function approximation in the feedback particle filter. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):1090–1117, 2020.
- Y. Tian, N. Panda, and Y. T. Lin. Liouville flow importance sampler. *arXiv:2405.06672*, 2024.
- N. G. Trillos, B. Hosseini, and D. Sanz-Alonso. From optimization to sampling through gradient flows. *Notices of the American Mathematical Society*, 70(6), 2023.
- F. Vargas, S. Padhy, D. Blessing, and N. Nüsken. Transport meets variational inference: Controlled Monte Carlo diffusions. In *The Twelfth International Conference on Learning Representations: ICLR 2024*, 2024.
- C. Villani. *Topics in optimal transportation*. American Mathematical Soc., 2003.
- C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
- G. Wahba and Y. Wang. *Representer Theorem*, pages 1–11. American Cancer Society, 2019. ISBN 9781118445112. doi:<https://doi.org/10.1002/9781118445112.stat08200>.
- L. Wang and N. Nüsken. Measure transport with kernel mean embeddings. *arXiv:2401.12967*, 2024.
- Y. Wang and W. Li. Information Newton’s flow: second-order optimization method in probability space. *arXiv:2001.04341*, 2020.
- J. Weidmann. *Linear operators in Hilbert spaces*, volume 68. Springer Science & Business Media, 2012.
- H. Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.
- Y. Yan, K. Wang, and P. Rigollet. Learning Gaussian mixtures using the Wasserstein-Fisher-Rao gradient flow. *arXiv:2301.01766*, 2023.
- L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.
- D.-X. Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of computational and Applied Mathematics*, 220(1-2):456–463, 2008.
- J.-J. Zhu and A. Mielke. Approximation, kernelization, and entropy-dissipation of gradient flows: from Wasserstein to Fisher-Rao. *preprint*, 2024. URL <https://jj-zhu.github.io/file/ZhuMielke24AppKerEntFR.pdf>.
- J. Zhuo, C. Liu, J. Shi, J. Zhu, N. Chen, and B. Zhang. Message passing Stein variational gradient descent. In *International Conference on Machine Learning (ICML)*, pages 6018–6027. PMLR, 2018.

## A Proofs for Sections 2 and 3

### A.1 Proof of Proposition 2

We present two proofs of Proposition 2, mainly because the notation and set up will be valuable for the proofs in Section B.

### A.1.1 Proof via the representer theorem

The first proof of Proposition 2 relies on the following version of the representer theorem, here recalled for convenience.

**Theorem 3** (Representer theorem, Wahba and Wang [2019]). *Let  $(H, \langle \cdot, \cdot \rangle_H)$  be a Hilbert space over  $\mathbb{R}$  and denote its continuous dual by  $(H', \langle \cdot, \cdot \rangle_{H'})$ . Let  $U' = \{u'_1, \dots, u'_N\} \subset H'$  be a collection of continuous linear functionals on  $H$ . Denote the set of associated Riesz representers by  $U = \{u_1, \dots, u_N\} \subset H$ , that is, we have that*

$$u'_j(v) = \langle u_j, v \rangle_H,$$

for all  $v \in H$  and  $j = 1, \dots, N$ . Furthermore, let  $\{y_1, \dots, y_N\} \subset \mathbb{R}$  be a collection of real numbers,  $\lambda > 0$  a regularisation parameter, and consider the regression problem

$$v^* \in \operatorname{argmin}_{v \in H} \left( \frac{1}{N} \sum_{j=1}^N (u'_j(v) - y_j)^2 + \lambda \|v\|_H^2 \right). \quad (37)$$

Then, (37) admits a unique solution  $v^*$ . Moreover,  $v^*$  belongs to the linear span of  $U$ , that is,

$$v^* = \sum_{i=1}^N \phi_i u_i,$$

for appropriate coefficients  $\phi_i \in \mathbb{R}$ . The coefficient vector  $(\phi_i)_{i=1}^N = \phi \in \mathbb{R}^N$  can be obtained as the unique solution to the linear system

$$(\frac{1}{N} \xi + \lambda I_{N \times N}) \phi = \mathbf{y},$$

where  $\mathbf{y} = (y_1, \dots, y_N)^\top \in \mathbb{R}^N$ , and the matrix  $\xi \in \mathbb{R}^{N \times N}$  is given by  $\xi_{ij} = \langle u_i, u_j \rangle_H$ .

*Proof of Proposition 2.* For  $j = 1, \dots, N$ , let us define the functionals  $u'_j : \mathcal{H}_k^d \rightarrow \mathbb{R}$  via

$$u'_j(v) := (S_\pi v)(X^j), \quad v \in \mathcal{H}_k^d.$$

Clearly, the functionals  $u'_j$  are linear, since  $S_\pi$  is a linear operator. Moreover, since  $k \in C^{1,1}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R})$ , we have that the divergence operators

$$\begin{aligned} \nabla \cdot |_{X^j} : \mathcal{H}_k^d &\rightarrow \mathbb{R}, \\ v &\mapsto (\nabla \cdot v)(X^j) = \sum_{l=1}^d \frac{\partial v^l(X^j)}{\partial x^l} \end{aligned}$$

are continuous, see Steinwart and Christmann [2008, Corollary 4.36]. Consequently, the functionals  $u'_j$  are continuous. We now claim that the corresponding Riesz representers are given by

$$u_j = k(\cdot, X^j) \nabla \log \pi(X^j) + \nabla_{X^j} k(\cdot, X^j) \in \mathcal{H}_k^d. \quad (38)$$

Indeed, by the reproducing property, as well as the derivative reproducing property [Zhou, 2008, Theorem 1] we have that

$$\langle u_j, v \rangle_{\mathcal{H}_k^d} = \nabla \log \pi(X^j) \cdot v(X^j) + (\nabla \cdot v)(X^j) = (S_\pi v)(X^j) = u'_j(v), \quad (39)$$

for all  $v \in \mathcal{H}_k^d$  and  $j = 1, \dots, N$ , as required. A direct computation using the reproducing and derivative reproducing properties now shows that

$$\langle u_i, u_j \rangle_{\mathcal{H}_k^d} = (\xi^{k, \nabla \log \pi})_{ij}, \quad i, j = 1, \dots, N,$$

where the right-hand side is defined according to (6) and (8). The claim now follows from Theorem 3 together with (38) and (39).  $\square$

### A.1.2 Proof via the Tikhonov regression formula

Here, we show that Problem 1 is a specific instance of Tikhonov-regularised least-squares problems. For fixed particle positions  $X_1, \dots, X_N \in \mathbb{R}^d$ , we start by equipping  $\mathbb{R}^N$  with the inner product

$$\langle x, y \rangle_N := \frac{1}{N} \sum_{i=1}^N x_i y_i, \quad x, y \in \mathbb{R}^N, \quad (40)$$

and define the linear operator  $S_{\pi,N} : (\mathcal{H}_k^d, \langle \cdot, \cdot \rangle_{\mathcal{H}_k^d}) \rightarrow (\mathbb{R}^N, \langle \cdot, \cdot \rangle_N)$  via

$$(S_{\pi,N}v)_i = (S_\pi v)(X^i), \quad i = 1, \dots, N, \quad v \in \mathcal{H}_k^d. \quad (41)$$

The objective in (5) can now be rewritten in the form

$$v^* \in \operatorname{argmin}_{v \in \mathcal{H}_k^d} \left( \|S_{\pi,N}v - \mathbf{h}_0\|_N^2 + \lambda \|v\|_{\mathcal{H}_k^d}^2 \right), \quad (42)$$

using the norm  $\|x\|^2 := \langle x, x \rangle_N$  in  $\mathbb{R}^N$ .

We now note that the objective in (42) coincides with the *Tikhonov functional* in Kirsch [2021, equation (2.12)] for the choices  $X = (\mathcal{H}_k^d, \langle \cdot, \cdot \rangle_{\mathcal{H}_k^d})$ ,  $Y = (\mathbb{R}^N, \langle \cdot, \cdot \rangle_N)$ ,  $K = S_{\rho,N}$ ,  $\alpha = \lambda$  and  $y = \mathbf{h}_0$ . By Kirsch [2021, Theorem 2.11],  $v_{N,\lambda}$  can be written in the form

$$v_{N,\lambda} = (\lambda I_{\mathcal{H}_k^d} + S_{\pi,N}^* S_{\pi,N})^{-1} S_{\pi,N}^* \mathbf{h}_0 \quad (43a)$$

$$= S_{\pi,N}^* (\lambda I_{N \times N} + S_{\pi,N} S_{\pi,N}^*)^{-1} \mathbf{h}_0, \quad (43b)$$

where  $I_{\mathcal{H}_k^d}$  denotes the identity operator on  $\mathcal{H}_k^d$ .

To apply the Tikhonov formulas (43), we seek an expression for the adjoint operator  $S_{\pi,N}^* : (\mathbb{R}^N, \langle \cdot, \cdot \rangle_N) \rightarrow (\mathcal{H}_k^d, \langle \cdot, \cdot \rangle_{\mathcal{H}_k^d})$ , characterised by

$$\langle c, S_{\pi,N}v \rangle_N = \langle S_{\pi,N}^* c, v \rangle_{\mathcal{H}_k^d}, \quad \text{for all } v \in \mathcal{H}_k^d, c \in \mathbb{R}^N.$$

A direct calculation using the reproducing and derivative reproducing properties [Zhou, 2008, Theorem 1] of  $k$  shows that

$$S_{\pi,N}^* c = \frac{1}{N} \sum_{j=1}^N (\nabla_{X^j} k(\cdot, X^j) + k(\cdot, X^j) \nabla \log \pi(X^j)) c_j, \quad c \in \mathbb{R}^N. \quad (44)$$

From this, we directly obtain that

$$(S_{\pi,N} S_{\pi,N}^* c)_i = \frac{1}{N} \sum_{j=1}^N \xi_{ij} c_j. \quad (45)$$

Combining (45) with (44), we see that (43b) coincides with the representation for  $v^*$  given in Proposition 1.

**Remark 11** (Weighted kernel ridge regression). For the construction in Section 4.2.1, it is crucial to slightly generalise the formulation in Problem 1,

$$v^* \in \operatorname{argmin}_{v \in \mathcal{H}_k^d} \left( \sum_{j=1}^N w^j ((S_\pi v)(X^j) - h_0(X^j))^2 + \lambda \|v\|_{\mathcal{H}_k^d}^2 \right), \quad (46)$$

replacing  $1/N$  by the particle weights  $w^j$ ; the motivation is to allow approximations of the form  $\pi \approx \sum_{i=1}^N w^i \delta_{X^i}$ . The Tikhonov regression approach can be adapted without difficulties: Instead of (40), we define the weight-dependent inner product

$$\langle x, y \rangle_w := \sum_{i=1}^N w_i x_i y_i, \quad x, y \in \mathbb{R}^N.$$

Proceeding analogously, we arrive at modifications of (44) and (45), with  $1/N$  replaced by  $w^j$ . The unique solution to (46) is therefore given by

$$v^* = \sum_{j=1}^N w^j \phi^j (k(\cdot, X^j) \nabla \log \pi(X^j) + \nabla_{X^j} k(\cdot, X^j)), \quad (47)$$

with  $(\phi^j)_{j=1}^N$  determined from

$$\sum_{j=1}^N (\xi^{k, \nabla \log \pi})_{ij} w^j \phi^j + \lambda \phi^i = h(X^i) - \sum_{j=1}^N w^j h(X^j), \quad i = 1, \dots, N. \quad (48)$$

Clearly, (47) and (48) generalise (9) and (10).

## A.2 Miscellaneous proofs

*Proof of Proposition 1.* We have

$$\partial_t \pi_t = -\frac{he^{-ht}\pi_0}{Z_t} - \frac{e^{-ht}\pi_0}{Z_t^2} \partial_t Z_t = -h\pi_t - \pi_t \frac{\partial_t Z_t}{Z_t}. \quad (49)$$

Moreover,

$$\frac{1}{Z_t} \partial_t Z_t = \frac{1}{Z_t} \partial_t \left( \int_{\mathbb{R}^d} e^{-ht} d\pi_0 \right) = -\frac{1}{Z_t} \int_{\mathbb{R}^d} he^{-ht} d\pi_0 = -\int_{\mathbb{R}^d} h d\pi_t, \quad (50)$$

where the exchange of differentiation and integration is permissible since  $he^{-h}$  is bounded, by the assumption that  $h$  is bounded from below. Combining (49) and (50), we see that the interpolation (1) satisfies

$$\partial_t \pi_t = -\pi_t \left( h - \int_{\mathbb{R}^d} h d\pi_t \right). \quad (51)$$

We now argue that the law associated to the ODE (4) is governed by the same equation (hence proving the claim by the well-posedness assumption). Indeed,  $\text{Law}(X_t)$  satisfies the continuity equation  $\partial_t \pi_t + \nabla \cdot (\pi_t v_t) = 0$ , see [Santambrogio, 2015, Section 4.1.2]. We now see that

$$\nabla \cdot (\pi_t v_t) = \rho_t S_{\pi_t} v_t = \pi_t \left( h - \int_{\mathbb{R}^d} h d\pi_t \right),$$

by (3), completing the proof.  $\square$

*Proof of Lemma 1.* First, we have that

$$\frac{d}{dt} (\nabla \log \rho_t(X_t)) = (\nabla \partial_t \log \rho_t)(X_t) + (\text{Hess} \log \rho_t(X_t)) \frac{dX_t}{dt}. \quad (52)$$

From  $\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0$ , see Santambrogio [2015], it follows that

$$\partial_t \log \rho_t = \frac{\partial_t \rho_t}{\rho_t} = -\frac{\nabla \cdot (\rho_t v_t)}{\rho_t} = -\nabla \cdot v_t - v_t \cdot \nabla \log \rho_t. \quad (53)$$

Plugging (53) into (52) and noticing that

$$\nabla(v_t \cdot \nabla \log \rho_t) = (\nabla v_t)(\nabla \log \rho_t)(X_t) + (\text{Hess} \log \rho_t(X_t)) v_t(X_t)$$

leads to the claimed identity.  $\square$

## B Proofs for Section 4

The objective of this section is to prove Propositions 3 and 5 as well as Theorems 1 and 2. We begin with the following simple estimate on the KSD-kernel  $\xi^{k, \nabla \log \pi}$ .

**Lemma 2** (Estimate on  $\xi^{k, \nabla \log \pi}$ ). *Under Assumption 1, there exists a constant  $C > 0$  such that*

$$|\xi^{k, \nabla \log \pi}(x, y)| \leq C(|\nabla \log \pi(x)| + |\nabla \log \pi(y)| + |\nabla \log \pi(x)| |\nabla \log \pi(y)|),$$

for all  $x, y \in \mathbb{R}^d$ .

*Proof.* This follows directly from the Cauchy-Schwarz and triangle inequalities, as well as from the boundedness of  $k$  and its derivatives.  $\square$

The proofs of Theorems 1 and 2 rely Tikhonov formulae of the form (43). The following lemma collects basic properties of the relevant operators.

**Lemma 3.** *Assume that  $\|\nabla \log \pi\|_{(L^2(\pi))^d} < \infty$  and that  $k$  is bounded, with bounded first-order derivatives. Then the following hold:*

1. *The Stein operator  $S_\pi$  is bounded from  $\mathcal{H}_k^d$  to  $L^2(\pi)$ , that is, there exists a constant  $C > 0$  such that*

$$\|S_\pi v\|_{L^2(\pi)} \leq C \|v\|_{\mathcal{H}_k^d},$$

*for all  $v \in \mathcal{H}_k^d$ .*

2. There exists a constant  $C > 0$  such that

$$\|\mathcal{T}_{k,\pi} \nabla \phi\|_{\mathcal{H}_k^d} \leq C \|\phi\|_{L^2(\pi)}, \quad (54)$$

for all  $\phi \in C_c^\infty(\mathbb{R}^d)$ . Therefore, there is a unique extension of  $\mathcal{T}_{k,\pi} \nabla$  to a bounded linear operator from  $L^2(\pi)$  to  $\mathcal{H}_k^d$  that we denote by the same symbol.

3. The adjoint of  $S_\pi : \mathcal{H}_k^d \rightarrow L^2(\pi)$  is given by  $-\mathcal{T}_{k,\pi} \nabla : L^2(\pi) \rightarrow \mathcal{H}_k^d$ , that is,

$$\langle S_\pi v, \phi \rangle_{L^2(\pi)} = -\langle v, \mathcal{T}_{k,\pi} \nabla \phi \rangle_{\mathcal{H}_k^d},$$

for all  $v \in \mathcal{H}_k^d$  and  $\phi \in L^2(\pi)$ .

*Proof.* 1.) There exists a constant  $C > 0$  such that

$$\|S_\pi v\|_{L^2(\pi)} \leq \|\nabla \log \pi\|_{(L^2(\rho))^d} \|v\|_{(L^2(\rho))^d} + \|\nabla \cdot v\|_{L^2(\pi)} \leq C \|v\|_{\mathcal{H}_k^d}, \quad (55)$$

for all  $v \in \mathcal{H}_k^d$ . The first inequality in (55) is implied by the triangle and Cauchy-Schwarz inequalities, while the second inequality follows from the regularity and boundedness assumptions on  $k$ , see Steinwart and Christmann [2008, Theorem 4.26 and Corollary 4.36].

2.) For  $\phi \in C_c^\infty(\mathbb{R}^d)$ , we have that

$$\|\mathcal{T}_{k,\pi} \nabla \phi\|_{\mathcal{H}_k^d}^2 = \left\langle \int_{\mathbb{R}^d} k(\cdot, y) \nabla \phi(y) \pi(dy), \int_{\mathbb{R}^d} k(\cdot, z) \nabla \phi(z) \pi(dz) \right\rangle_{\mathcal{H}_k^d} \quad (56a)$$

$$= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \nabla \phi(y) \cdot k(y, z) \nabla \phi(z) \pi(dy) \pi(dz) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \xi^{k, \nabla \log \pi}(y, z) \phi(y) \phi(z) \pi(dy) \pi(dz). \quad (56b)$$

From (56a) to (56b) we have used the fact that Bochner integration commutes with bounded linear operators [Steinwart and Christmann, 2008, equation (A.32)] to change the order of integration and inner products and apply the reproducing property. The second identity in (56b) follows from integration by parts. We now obtain the bound (54) from Lemma 2 and the fact that  $\|\nabla \log \pi\|_{(L^2(\pi))^d} < \infty$  by assumption.

3.) We have

$$-\langle \mathcal{T}_{k,\pi} \nabla \phi, v \rangle_{\mathcal{H}_k^d} = -\int_{\mathbb{R}^d} \langle k(\cdot, y) \nabla \phi(y), v \rangle_{\mathcal{H}_k^d} \rho(dy) = -\int_{\mathbb{R}^d} \nabla \phi \cdot v \, d\pi = \langle \phi, S_\pi v \rangle_{L^2(\pi)}, \quad (57)$$

as required, for all  $v \in \mathcal{H}_k^d$  and  $\phi \in L^2(\pi)$ . As in the proof of the second statement, we have made use of the fact that Bochner integration and bounded linear operators commute.  $\square$

Proposition 3 can now be obtained from a Tikhonov-regularised least-squares formulation (cf. the proof of Proposition 2 in Appendix A.1.2):

*Proof of Proposition 3.* As in the proof of Proposition 2, we can reformulate (14) as

$$v_\infty^* \in \operatorname{argmin}_{v \in \mathcal{H}_k^d} \left( \|S_\pi v - h_{0,\infty}\|_{L^2(\pi)}^2 + \lambda \|v\|_{\mathcal{H}_k^d}^2 \right).$$

Building on Lemma 3 and Kirsch [2021, Theorem 2.11], there exists a unique minimiser, given by

$$v_\infty^* = S_\pi^*(\lambda I_{L^2(\pi)} + S_\pi S_\pi^*)^{-1} h_{0,\infty} \quad (58a)$$

$$= (\lambda I_{\mathcal{H}_k^d} + S_\pi^* S_\pi)^{-1} S_\pi^* h_{0,\infty}. \quad (58b)$$

The claim now follows from  $S_\pi^* = -\mathcal{T}_{k,\pi} \nabla$ , see Lemma 3, and from the fact that the equation

$$(S_\pi S_\pi^* + \lambda I_{L^2(\pi)})\phi = h_{0,\infty}$$

can be written in the form (16), multiplying both sides by  $\pi$ .  $\square$

*Proof of Proposition 5.* Using the Stein equation  $S_\pi v_\infty = h - \int_{\mathbb{R}^d} h \, d\pi$ , the fact that  $\frac{1}{N} \sum_{i=1}^N = \int_{\mathbb{R}^d} h \, d\pi$ , and (43b), we can write

$$v_{N,0} = S_{\pi,N}^* (S_{\pi,N} S_{\pi,N}^*)^{-1} S_{\pi,N} v_\infty.$$

Notice that the operator  $S_{\pi,N} S_{\pi,N}^* : \mathbb{R}^N \rightarrow \mathbb{R}^N$  is invertible since  $\xi \in \mathbb{R}^{N \times N}$  is invertible, see equation (45). The operator  $P_X := S_{\pi,N}^* (S_{\pi,N} S_{\pi,N}^*)^{-1} S_{\pi,N}$  is an orthogonal projection onto the subspace of  $\mathcal{H}_k^d$  defined in (25). Indeed, it is immediate that  $P_X$  is self-adjoint (and positive definite), and that  $P_X^2 = P_X$ . We also have  $\operatorname{Ran} P_X = \operatorname{Ran} S_{\pi,N}^*$  which coincides with the subspace (25), see equation (44).  $\square$

Before proceeding to the proof of Theorem 1, we need the following technical lemma, which is a slight extension of Villani [2003, Theorem 7.12] to the context of Hilbert spaces and Bochner integrals.

**Lemma 4.** *Let  $H$  be a separable Hilbert space with corresponding norm  $\|\cdot\|_H$ . Let  $\phi : \mathbb{R}^d \rightarrow H$  be Borel-measurable. Assume that there exist constants  $C > 0$  and  $p > 0$  such that*

$$\|\phi(x)\|_H \leq C(1 + |x|^p), \quad (59)$$

for all  $x \in \mathbb{R}^d$ . Let  $\mu_k \subset \mathcal{P}(\mathbb{R}^d)$  be a sequence of probability measures with finite  $p^{\text{th}}$  moments that converges in  $W^p$  to some  $\mu \in \mathcal{P}(\mathbb{R}^d)$ . Then

$$\int_{\mathbb{R}^d} \phi \, d\mu_k \rightarrow \int_{\mathbb{R}^d} \phi \, d\mu$$

as Bochner integrals in  $H$ .

*Proof.* By Villani [2003, Theorem 7.12], convergence in  $W^p$  implies

$$\lim_{R \rightarrow \infty} \limsup_{k \rightarrow \infty} \int_{|x| \geq R} |x|^p \, d\mu_k(x) = 0. \quad (60)$$

For arbitrary  $R > 0$ , we have

$$\left\| \int_{\mathbb{R}^d} \phi \, d\mu_k - \int_{\mathbb{R}^d} \phi \, d\mu \right\|_H \leq \left\| \int_{|x| < R} \phi \, d\mu_k - \int_{|x| < R} \phi \, d\mu \right\|_H \quad (61a)$$

$$+ C \left( \int_{|x| \geq R} (1 + |x|^p) \, d\mu_k + \int_{|x| \geq R} (1 + |x|^p) \, d\mu \right). \quad (61b)$$

By Hable and Christmann [2011, Theorem 5.1], the right-hand side of (61a) vanishes in the limit as  $k \rightarrow \infty$ . Therefore, we see that

$$\limsup_{k \rightarrow \infty} \left\| \int_{\mathbb{R}^d} \phi \, d\mu_k - \int_{\mathbb{R}^d} \phi \, d\mu \right\|_H \leq C \limsup_{k \rightarrow \infty} \left( \int_{|x| \geq R} (1 + |x|^p) \, d\mu_k + \int_{|x| \geq R} (1 + |x|^p) \, d\mu \right).$$

The result now follows from (60) by taking the limit as  $R \rightarrow \infty$ .  $\square$

*Proof of Theorem 1.* We follow a similar calculation by Smale and Zhou [2007, Section 3] and write

$$\begin{aligned} v_{N,\lambda} - v_{\infty,\lambda} &= (\lambda I_{\mathcal{H}_k^d} + S_{\pi,N}^* S_{\pi,N})^{-1} \left( S_{\pi,N}^* h_{0,N} - (\lambda I_{\mathcal{H}_k^d} + S_{\pi,N}^* S_{\pi,N}) v_{\infty,\lambda} \right) \\ &= (\lambda I_{\mathcal{H}_k^d} + S_{\pi,N}^* S_{\pi,N})^{-1} (S_{\pi,N}^*(h_{0,N} - S_{\pi,N} v_{\infty,\lambda}) - \lambda v_{\infty,\lambda}) \\ &= (\lambda I_{\mathcal{H}_k^d} + S_{\pi,N}^* S_{\pi,N})^{-1} (S_{\pi,N}^*(h_{0,N} - S_{\pi,N} v_{\infty,\lambda}) - (S_{\pi}^* h_{0,\infty} - S_{\pi}^* S_{\pi} v_{\infty,\lambda})) \\ &= (\lambda I_{\mathcal{H}_k^d} + S_{\pi,N}^* S_{\pi,N})^{-1} ((S_{\pi}^* S_{\pi} - S_{\pi,N}^* S_{\pi,N}) v_{\infty,\lambda} + S_{\pi,N}^* h_{0,N} - S_{\pi}^* h_{0,\infty}), \end{aligned}$$

using the sample Stein operator  $S_{\pi,N}$  defined in (41), as well as the Tikhonov formulae (43) and (58). From this, we see that

$$\|v_{N,\lambda} - v_{\infty,\lambda}\|_{\mathcal{H}_k^d} \leq \frac{1}{\lambda} \left( \| (S_{\pi}^* S_{\pi} - S_{\pi,N}^* S_{\pi,N}) v_{\infty,\lambda} \|_{\mathcal{H}_k^d} + \| S_{\pi,N}^* h_{0,N} - S_{\pi}^* h_{0,\infty} \|_{\mathcal{H}_k^d} \right). \quad (63)$$

To estimate the first term on the right-hand side of (63), we introduce the notation  $\pi^{(N)} := \frac{1}{N} \sum_{i=1}^N \delta_{X^i}$  and compute

$$\begin{aligned} S_{\pi,N}^* S_{\pi,N} v_{\infty,\lambda} &= \frac{1}{N} \sum_{i=1}^N \left[ (\nabla_{X^i} k(\cdot, X^i) + k(\cdot, X^i) \nabla \log \pi(X^i)) (S_{\pi} v_{\infty,\lambda})(X^i) \right] \\ &= \int_{\mathbb{R}^d} (\nabla_y k(\cdot, y) + k(\cdot, y) \nabla \log \pi(y)) (S_{\pi} v_{\infty,\lambda})(y) \pi^{(N)}(dy), \end{aligned}$$

using (41) and (44). Similarly,

$$S_{\pi}^* S_{\pi} v_{\infty,\lambda} = \int_{\mathbb{R}^d} (\nabla_y k(\cdot, y) + k(\cdot, y) \nabla \log \pi(y)) (S_{\pi} v_{\infty,\lambda})(y) \pi(dy).$$

Consequently,

$$\|(S_\pi^* S_\pi - S_{\pi,N}^* S_{\pi,N}) v_{\infty,\lambda}\|_{\mathcal{H}_k^d}^2 = \left\| \int_{\mathbb{R}^d} (\nabla_y k(\cdot, y) + k(\cdot, y) \nabla \log \pi(y)) (S_\pi v_{\infty,\lambda})(y) (\pi - \pi^{(N)})(dy) \right\|_{\mathcal{H}_k^d}^2.$$

To show convergence of the left-hand side, by Lemma 4 it is sufficient to show that the map  $\phi : \mathbb{R}^d \ni y \mapsto (\nabla_y k(\cdot, y) + k(\cdot, y) \nabla \log \pi(y)) (S_\pi v_{\infty,\lambda})(y) \in \mathcal{H}_k^d$  satisfies the growth bound (59). Indeed, we have the following bound,

$$\|(\nabla_y k(\cdot, y) + k(\cdot, y) \nabla \log \pi(y)) (S_\pi v_{\infty,\lambda})(y)\|_{\mathcal{H}_k^d} \leq |S_\pi v_{\infty,\lambda}(y)| \|\nabla_y k(\cdot, y) + k(\cdot, y) \nabla \log \pi(y)\|_{\mathcal{H}_k^d} \quad (65a)$$

$$= |S_\pi v_{\infty,\lambda}(y)| \sqrt{\xi(y, y)} \leq \tilde{C}_1(\lambda) (1 + |\nabla \log \pi(y)|^2) \leq \tilde{C}_2(\lambda) (1 + |y|^p), \quad (65b)$$

for some constants  $\tilde{C}_1, \tilde{C}_2 > 0$  that might depend on  $\lambda$ . Here, we have used the fact that  $v_{\infty,\lambda}$  and its first derivatives are bounded (since  $v_{\infty,\lambda} \in \mathcal{H}_k^d$  and  $k$  and its first derivatives are bounded), as well as the growth assumption on  $\nabla \log \pi$  and the estimate from Lemma 2.

We conclude the proof by providing a similar estimate to show convergence of the second term on the right-hand side of (63). We have

$$\|S_{\pi,N}^* h_{0,N} - S_\pi^* h_{0,\infty}\|_{\mathcal{H}_k^d} = \left\| S_{\pi,N}^* \left( h - \frac{1}{N} \sum_{i=1}^N h(X^i) \right) - S_\pi^* \left( h - \int_{\mathbb{R}^d} h d\pi \right) \right\|_{\mathcal{H}_k^d} \quad (66a)$$

$$\leq \| (S_{\pi,N}^* - S_\pi^*) h \|_{\mathcal{H}_k^d} + \left\| S_{\pi,N}^* \left( \frac{1}{N} \sum_{i=1}^N h(X^i) - \int_{\mathbb{R}^d} h d\pi \right) \right\|_{\mathcal{H}_k^d} + \left\| (S_{\pi,N}^* - S_\pi^*) \int_{\mathbb{R}^d} h d\pi \right\|_{\mathcal{H}_k^d} \quad (66b)$$

$$\leq \left\| \int_{\mathbb{R}^d} (\nabla_y k(\cdot, y) + k(\cdot, y) \nabla \log \pi(y)) h(y) (\pi - \pi^{(N)})(dy) \right\|_{\mathcal{H}_k^d} \quad (66c)$$

$$+ \left\| \left( \int_{\mathbb{R}^d} (\nabla_y k(\cdot, y) + k(\cdot, y) \nabla \log \pi(y)) (\pi - \pi^{(N)})(dy) \right) \left( \int_{\mathbb{R}^d} h d\pi - \frac{1}{N} \sum_{i=1}^N h(X^i) \right) \right\|_{\mathcal{H}_k^d} \quad (66d)$$

For the first term in (66b), notice that

$$\| (S_{\pi,N}^* - S_\pi^*) h \|_{\mathcal{H}_k^d} = \left\| \int_{\mathbb{R}^d} (\nabla_y k(\cdot, y) + k(\cdot, y) \nabla \log \pi(y)) h(y) (\pi - \pi^{(N)})(dy) \right\|_{\mathcal{H}_k^d},$$

with

$$\|(\nabla_y k(\cdot, y) + k(\cdot, y) \nabla \log \pi(y)) h(y)\|_{\mathcal{H}_k^d} \leq \sqrt{\xi(y, y)} |h(y)| \leq \tilde{C}(1 + |y|^p),$$

as in (65), with a possibly different constant  $\tilde{C}$ . The second and the third term in (66b) can be bounded in the same way, and we omit the details. Putting everything together, convergence to zero of (66a) follows from Lemma 4.  $\square$

As a preparation for the proof of Theorem 2, we need the following lemma, characterising the kernel of  $S_\pi S_\pi^*$ .

**Lemma 5.** *Let Assumptions 1 and 2 be satisfied. Then  $\ker S_\pi S_\pi^*$  consists of constants.*

*Proof.* Let us first introduce the linear subspace  $L_0^2(\pi) \subset L^2(\pi)$ ,

$$L_0^2(\pi) = \left\{ \phi \in L^2(\pi) : \int_{\mathbb{R}^d} \phi d\pi = 0 \right\}, \quad (67)$$

equipped with the restriction of the  $L^2(\pi)$ -inner product. Since  $S_\pi^* = -\mathcal{T}_{k,\pi} \nabla$  vanishes on constants, we may view  $S_\pi S_\pi^*$  as a bounded linear operator on  $L_0^2(\pi)$  and show that  $\ker S_\pi S_\pi^* = \{0\}$ . To that end, notice that  $\ker S_\pi S_\pi^* = \ker S_\pi^* = (\text{Ran } S_\pi)^\perp$ , where the orthogonal complement is taken with respect to the  $L^2(\pi)$  inner product, see Weidmann [2012, Theorem 4.13, b)]. It is thus sufficient to show that if for some  $\phi \in C_c^\infty(\mathbb{R}^d) \cap L_0^2(\pi)$ , we have

$$\langle \phi, S_\pi v \rangle_{L^2(\pi)} = 0, \quad \text{for all } v \in \mathcal{H}_k^d, \quad (68)$$

then necessarily  $\phi = 0$ . For this, first notice that

$$\langle \phi, S_\pi v \rangle_{L^2(\pi)} = - \int_{\mathbb{R}^d} \nabla \phi \cdot v d\pi.$$

Assume now that  $\phi \in C_c^\infty(\mathbb{R}^d) \cap L_0^2(\pi)$  is fixed such that (68) holds. By density of  $\mathcal{H}_k^d$  in  $(L^2(\pi))^d$ , we can choose a sequence  $(v_n) \subset \mathcal{H}_k^d$  such that  $v_n \rightarrow \nabla\phi$  in  $(L^2(\pi))^d$ . We then obtain

$$0 = \langle \phi, S_\pi v_n \rangle_{L^2(\pi)} = - \int_{\mathbb{R}^d} \nabla\phi \cdot v_n \, d\pi \rightarrow - \int_{\mathbb{R}^d} |\nabla\phi|^2 \, d\pi$$

by continuity, which clearly implies  $\phi = 0$ .  $\square$

We also need the following compactness result:

**Lemma 6.** *Let Assumptions 1 and 2 be satisfied. Then  $S_\pi S_\pi^* : L^2(\pi) \rightarrow L^2(\pi)$  is compact.*

*Proof.* Using Lemma 3, first notice that

$$\begin{aligned} S_\pi S_\pi^* \phi &= -S_\pi \mathcal{T}_{k,\pi} \nabla\phi = -S_\pi \int_{\mathbb{R}^d} k(\cdot, y) \nabla\phi(y) \pi(dy) \\ &= S_\pi \int_{\mathbb{R}^d} (\nabla_y k(\cdot, y) + k(\cdot, y) \nabla \log \pi(y)) \phi(y) \pi(dy) = \int_{\mathbb{R}^d} \xi^{k,\nabla \log \pi}(\cdot, y) \phi(y) \pi(dy). \end{aligned} \quad (69)$$

From Lemma 2 we now have

$$\int_{\mathbb{R}^d} \xi^{k,\nabla \log \pi}(x, x) \pi(dx) < \infty.$$

The statement therefore follows from Steinwart and Christmann [2008, Theorem 4.27].  $\square$

*Proof of Theorem 2.* We begin with the estimate

$$\|S_\pi v_{N,\lambda} - h_{0,\infty}\|_{L^2(\pi)} \leq \|S_\pi(v_{N,\lambda} - v_{\infty,\lambda})\|_{L^2(\pi)} + \|S_\pi v_{\infty,\lambda} - h_{0,\infty}\|_{L^2(\pi)} \quad (70a)$$

$$\leq \|S_\pi\|_{\mathcal{H}_k^d \rightarrow L^2(\pi)} \|v_{N,\lambda} - v_{\infty,\lambda}\|_{\mathcal{H}_k^d} + \|S_\pi v_{\infty,\lambda} - h_{0,\infty}\|_{L^2(\pi)}. \quad (70b)$$

Notice that the operator norm  $\|S_\pi\|_{\mathcal{H}_k^d \rightarrow L^2(\pi)}$  is finite by Lemma 3, and that for fixed  $\lambda > 0$ , the term  $\|v_{N,\lambda} - v_{\infty,\lambda}\|_{\mathcal{H}_k^d}$  converges to zero as  $N \rightarrow \infty$ , as a consequence of Theorem 2.

Note that the second term in (70b) does not depend on  $N$ . We show that it converges to zero as  $\lambda \rightarrow 0$ . Recall the linear subspace  $L_0^2(\pi) \subset L^2(\pi)$  defined in (67). Clearly,  $h_{0,\infty} \in L_0^2(\pi)$ , and by Lemmas 5 and 6,  $S_\pi S_\pi^*$  acts as a self-adjoint, strictly positive definite, and compact operator on  $L_0^2(\pi)$ . Consequently, there exists an orthonormal basis  $(e_i)_{i=1}^\infty$  in  $L_0^2(\pi)$  such that  $S_\pi S_\pi^* e_i = \mu_i e_i$ , with strictly positive eigenvalues  $\mu_i$ . Writing  $h_{0,\infty} = \sum_{i=1}^\infty h_i e_i$ , we see from (58a) that

$$\|S_\pi v_{\infty,\lambda} - h_{0,\infty}\|_{L^2(\pi)}^2 = \sum_{i=1}^\infty \left( \frac{\mu_i}{\lambda + \mu_i} - 1 \right)^2 h_i^2. \quad (71)$$

Since the term in parenthesis is bounded by one, the dominated convergence theorem allows us to take the limit as  $\lambda \rightarrow 0$  termwise. As these limits are zero, we conclude that  $\|S_\pi v_{\infty,\lambda} - h_{0,\infty}\|_{L^2(\pi)}^2 \rightarrow 0$ . Note that the fact that  $\mu_i > 0$  is crucial, since otherwise the corresponding term would not converge to zero.

We now conclude as follows: For given  $\varepsilon > 0$ , we can choose  $\lambda > 0$  such that  $\|S_\pi v_{\infty,\lambda} - h_{0,\infty}\|_{L^2(\pi)} < \frac{\varepsilon}{2}$ . For this fixed value of  $\lambda$ , we have from Theorem 1 that  $\|S_\pi\|_{\mathcal{H}_k^d \rightarrow L^2(\pi)} \|v_{N,\lambda} - v_{\infty,\lambda}\|_{\mathcal{H}_k^d}$  converges to zero as  $N \rightarrow \infty$ . We can thus take  $N_0 \in \mathcal{N}$  such that this term is bounded from above by  $\frac{\varepsilon}{2}$ , for all  $N \geq N_0$ .  $\square$

*Proof of Proposition 4.* The Stein-Poisson equation (16) implies

$$\lambda \phi^{(\lambda)} = h - \int_{\mathbb{R}^d} h \, d\pi + \frac{1}{\pi} \nabla \cdot (\pi \mathcal{T}_{k,\pi} \nabla\phi^{(\lambda)}) = h_{0,\infty} - S_\pi v_{\infty,\lambda},$$

using the fact that  $v_{\infty,\lambda} = -\mathcal{T}_{k,\pi} \nabla\phi^{(\lambda)}$ . The first claim now follows from (71), together with reasoning following it in the proof of Theorem 2. For the second claim, we follow an argument from De Vito et al. [2005, Appendix A]: Using  $\mu_i/(\lambda + \mu_i) - 1 = -1/(1 + \mu_i/\lambda)$ , we estimate (71),

$$\|S_\pi v_{\infty,\lambda} - h_{0,\infty}\|_{L^2(\pi)}^2 = \sum_{i=1}^\infty \left( \frac{1}{1 + \frac{\mu_i}{\lambda}} \right)^2 h_i^2 \leq \sum_{i=1}^\infty \left( \frac{1}{\left( \frac{\mu_i}{\lambda} \right)^\alpha} \right)^2 h_i^2 = \lambda^{2\alpha} \sum_{i=1}^\infty \mu_i^{-2\alpha} h_i^2, \quad (72)$$

making use of the inequality  $x^\alpha \leq x + 1$ , which holds for all  $\alpha \in (0, 1]$  and  $x \in [0, \infty)$ . From the integral representation of  $S_\pi S_\pi^*$  in (69), we see that the RKHS associated to  $\xi^{k,\nabla \log \pi}$  is isomorphic to the range  $(S_\pi S_\pi^*)^{1/2} L^2(\pi)$ , see also Steinwart and Christmann [2008, Theorem 4.51]. By the definition of the fractional powers  $\mathcal{H}_{k,\nabla \log \pi}^\alpha$  [Muandet et al., 2017, Definition 4.11], the right-most expression is finite if  $h_{0,\infty} \in \mathcal{H}_{k,\nabla \log \pi}^\alpha$ , and so the claim follows.  $\square$

## C Proof of Proposition 6

*Proof of Proposition 6.* We first summarise the main steps of the proof. For fixed  $t \in (-\varepsilon, \varepsilon)$ , let us denote by  $(v_\tau^t)_{\tau \in [0,1]}$  the family of vector fields that realise the Stein optimal transport between  $\pi_0$  and  $\pi_t$ , in the sense of (28), noticing that  $t$  is replaced by  $\tau$  in that equation. The Stein optimal transport maps then take the form

$$F_t(x) = x + \int_0^1 v_\tau^t(X_\tau^{t,x}) d\tau,$$

where  $X_\tau^t$  solves the ODE

$$\frac{dX_\tau^{t,x}}{d\tau} = v_\tau^t(X_\tau^{t,x}), \quad X_\tau^{t,x} = x.$$

We then proceed as follows: Firstly, we show that there exists a sequence  $t_n \rightarrow 0$  with  $t_n > 0$  such that

$$\frac{F_{t_n}(x) - x}{t_n} = \frac{1}{t_n} \int_0^{t_n} v_\tau^{t_n}(X_\tau^{t_n,x}) d\tau \xrightarrow{n \rightarrow \infty} v^*(x), \quad (73)$$

for some  $v^* \in \mathcal{H}_k^d$  and all  $x \in \mathbb{R}^d$ . Secondly, we show that

$$\partial_t \pi_t|_{t=0} + \nabla \cdot (\pi_0 v^*) = 0, \quad (74)$$

in the weak sense:  $v^*$  is compatible with the dynamics of the curve  $(\pi_t)_{t \in (-\varepsilon, \varepsilon)}$ . Next, we show that  $v^*$  is minimal among vector fields that satisfy (74), in the sense of the  $\|\cdot\|_{\mathcal{H}_k^d}$ -norm. Lastly, we show that these characterisations of  $v^*$  allow us to conclude convergence along any sequence  $t_n$  in (73), to the same limit, and that this limit is characterised by the Stein-Poisson equation (31).

*Step 1.* We first argue that

$$\|v_\tau^t\|_{\mathcal{H}_k^d} = d_k(\pi_0, \pi_t), \quad (75)$$

for all  $t \in (-\varepsilon, \varepsilon)$  and  $\tau \in [0, 1]$ . In other words, minimisers in (26) are automatically parameterised by arc-length, as  $\|v_\tau^t\|_{\mathcal{H}_k^d}$  is constant in  $\tau \in [0, 1]$ . This is a standard fact for geodesic distances on Riemannian manifolds of the form (26), see, for instance, [Jost and Jost \[2008, Section 1.4\]](#). For convenience, let us repeat the argument:

1. Firstly, Jensen's inequality implies the energy-length comparison

$$\int_0^1 \|v_\tau^t\|_{\mathcal{H}_k^d}^2 d\tau \geq \left( \int_0^1 \|v_\tau^t\|_{\mathcal{H}_k^d} d\tau \right)^2, \quad (76)$$

with equality if and only if  $\|v_\tau^t\|_{\mathcal{H}_k^d}$  is constant in  $\tau$ .

2. Secondly, the right-hand side of (76) is reparameterisation-invariant in the following sense: If  $\gamma : [0, 1] \rightarrow [0, 1]$  is a reparameterisation (meaning that  $\gamma$  is absolutely continuous on  $(0, 1)$ , with  $\gamma(0) = 0$ ,  $\gamma(1) = 1$  and  $\gamma' > 0$  Lebesgue almost everywhere on  $(0, 1)$ ), then the reparameterised vector field  $v_\tau^t$  (defined by  $\tilde{v}_\tau^t := \gamma'(\tau)v_{\gamma(\tau)}^t$  on  $\tau \in (0)$  and  $\tilde{v}_0^t = v_0^t$  as well as  $\tilde{v}_1^t = v_1^t$ ) satisfies

$$\int_0^1 \|v_\tau^t\|_{\mathcal{H}_k^d} d\tau = \int_0^1 \|\tilde{v}_\tau^t\|_{\mathcal{H}_k^d} d\tau. \quad (77)$$

Combining (76) and (77), it follows that any given  $(v_\tau^t)_{\tau \in [0,1]}$  can be reparameterised in such a way that (75) holds, without affecting the right-hand side of (76), but minimising the right-hand side.

From (75), the fact that  $v_\tau^0 = 0$ , and Lipschitz continuity of  $t \mapsto d_k(\pi_0, \pi_t)$ , it follows that the set

$$\left\{ \frac{1}{t} \int_0^1 v_\tau^t d\tau : t \in (-\varepsilon, \varepsilon) \setminus \{0\} \right\} \quad (78)$$

is bounded in  $\mathcal{H}_k^d$ . As a consequence, the Banach-Alaoglu theorem allows us to extract a sequence  $\frac{1}{t_n} \int_0^1 v_\tau^{t_n} d\tau$  with  $t_n \rightarrow 0$  that converges weakly to some  $v^* \in \mathcal{H}_k^d$ ; that is

$$\left\langle \frac{1}{t_n} \int_0^1 v_\tau^{t_n} d\tau, h \right\rangle_{\mathcal{H}_k^d} \xrightarrow{n \rightarrow \infty} \langle v^*, h \rangle_{\mathcal{H}_k^d},$$

for all  $h \in \mathcal{H}_k^d$ . By choosing  $h$  appropriately and using the reproducing property (and the fact that bounded linear operators commute with Bochner integration), we see that  $\frac{1}{t_n} \int_0^1 v_\tau^{t_n} d\tau$  converges to  $v^*$  pointwise (for example, choosing  $h = (k(x, \cdot), 0 \dots, 0)$ , with arbitrary  $x \in \mathbb{R}^d$ , shows that the first components converge).

We now show the convergence in (73), for fixed  $x \in \mathbb{R}^d$ . To that end, notice that

$$\left| \frac{F_{t_n}(x) - x}{t_n} - v^*(x) \right| \leq \left| \frac{1}{t_n} \int_0^1 (v_\tau^{t_n}(X_\tau^{t_n, x}) - v_\tau^{t_n}(x)) d\tau \right| + \left| \frac{1}{t_n} \int_0^1 v_\tau^{t_n}(x) d\tau - v^*(x) \right|. \quad (79)$$

According to the previous arguments, the second term on the right-hand side converges to zero. The first term can be bounded from above by a constant times

$$\frac{1}{t_n} \int_0^1 \|v_\tau^{t_n}\|_{\mathcal{H}_k^d} |X_\tau^{t_n, x} - x| d\tau, \quad (80)$$

owing to the fact that  $v_\tau^{t_n}$  is Lipschitz continuous (since  $k$  has bounded derivatives by assumption), and the Lipschitz constant is controlled by the RKHS-norm [Steinwart and Christmann, 2008, Corollary 4.36]. Since  $\frac{1}{t_n} \|v_\tau^{t_n}\|_{\mathcal{H}_k^d}$  is bounded (again, because of (75) and the Lipschitz property of  $t \mapsto d_k(\pi_0, \pi_t)$ ) and

$$|X_\tau^{t_n, x} - x| \leq \int_0^\tau |v_s^{t_n}(X_s^{t_n, x})| ds \lesssim \int_0^\tau \|v_s^{t_n}(X_s^{t_n, x})\|_{\mathcal{H}_k^d} ds = \tau d_k(\pi_0, \pi_{t_n}) \xrightarrow{n \rightarrow \infty} 0,$$

we conclude that (80) converges to zero, and therefore (73) holds.

*Step 2.* Let us show that any  $v^*$  identified in Step 1 satisfies the continuity equation (74). For  $\psi \in C_c^\infty(\mathbb{R}^d)$ , we have

$$\frac{1}{t_n} \left( \int_{\mathbb{R}^d} \psi d\pi_{t_n} - \int_{\mathbb{R}^d} \psi d\pi_0 \right) = \frac{1}{t_n} \left( \int_{\mathbb{R}^d} \psi d((F_{t_n})_\# \pi_0) - \int_{\mathbb{R}^d} \psi d\pi_0 \right) = \frac{1}{t_n} \int_{\mathbb{R}^d} (\psi(F_{t_n}(x)) - \psi(x)) d\pi_0.$$

The left-hand side of this equality converges to  $\int_{\mathbb{R}^d} \psi \partial_t \pi|_{t=0} dx$ , while the left-hand side converges to  $\int_{\mathbb{R}^d} \nabla \psi \cdot v^* d\pi_0$ , using (73).

*Step 3.* In this step, we establish that the weak limit  $v^*$  is minimal among solutions to (74), in the sense of  $\|\cdot\|_{\mathcal{H}_k^d}$ . To achieve this, we compare optimal Stein transport interpolations encoded by the vector fields  $v_\tau^t$  to the (suboptimal) flow of  $(\pi_t)_{t \in (-\varepsilon, \varepsilon)}$ . For  $t \in (-\varepsilon, \varepsilon)$  and  $\tau \in [0, 1]$ , we set

$$w_\tau^t := t \mathcal{T}_{k, \pi_{\tau t}} \nabla \phi_{\tau t}, \quad \text{and} \quad \rho_\tau := \pi_{\tau t}.$$

Notice that  $\rho_0 = \pi_0$  and  $\rho_1 = \pi_t$ , as well as  $\partial_\tau \rho_\tau + \nabla \cdot (\rho_\tau w_\tau^t) = 0$ , and therefore  $(w_\tau^t)_{\tau \in [0, 1]}$  is a (suboptimal) competitor for  $(v_\tau^t)_{\tau \in [0, 1]}$  in the transport formulation (26) for  $d_k(\pi_0, \pi_t)$ . We therefore have

$$\begin{aligned} \left\| \frac{1}{t_n} \int_0^1 v_\tau^{t_n} d\tau \right\|_{\mathcal{H}_k^d}^2 &\leq \frac{1}{t_n^2} \int_0^1 \|v_\tau^{t_n}\|_{\mathcal{H}_k^d}^2 d\tau \leq \frac{1}{t_n^2} \int_0^1 \|w_\tau^{t_n}\|_{\mathcal{H}_k^d}^2 d\tau = \int_0^1 \|\mathcal{T}_{k, \pi_{\tau t_n}} \nabla \phi_{\tau t_n}\|_{\mathcal{H}_k^d}^2 d\tau \\ &\xrightarrow{n \rightarrow \infty} \|\mathcal{T}_{k, \pi_0} \nabla \phi_0\|_{\mathcal{H}_k^d}^2, \end{aligned}$$

implying that  $\|v^*\|_{\mathcal{H}_k^d} \leq \|\mathcal{T}_{k, \pi_0} \nabla \phi_0\|_{\mathcal{H}_k^d}$  by the weak lower semicontinuity of  $\|\cdot\|_{\mathcal{H}_k^d}$ .

*Step 4.* The fact that  $v^*$  satisfies (74) and  $\|v^*\|_{\mathcal{H}_k^d} \leq \|\mathcal{T}_{k, \pi_0} \nabla \phi_0\|_{\mathcal{H}_k^d}$  implies  $v^* = \mathcal{T}_{k, \pi_0} \nabla \phi_0$ , because solutions to the continuity equation with minimal RKHS-norm take precisely this form [Duncan et al., 2023, Proposition 5]. In this step, we make this argument precise. The Helmholtz decomposition in  $\mathcal{H}_k^d$  [Duncan et al., 2023, Proposition 6] is

$$\mathcal{H}_k^d = (L_{\text{div}}^2(\pi_0) \cap \mathcal{H}_k^d) \oplus \overline{\mathcal{T}_{k, \pi_0} \nabla C_c^\infty(\mathbb{R}^d)}^{\mathcal{H}_k^d}, \quad (82)$$

where the two subspaces are orthogonal in  $\mathcal{H}_k^d$ , and  $L_{\text{div}}^2(\pi_0)$  is the space weighted divergence-free vector fields,

$$L_{\text{div}}^2(\pi_0) = \{v \in (L^2(\pi_0))^d : \langle v, \nabla \phi \rangle_{(L^2(\pi_0))^d}, \text{ for all } \phi \in C_c^\infty(\mathbb{R}^d)\}.$$

Because both  $v^*$  and  $w := \mathcal{T}_{k, \pi_0} \nabla \phi_0$  satisfy the continuity equation (74), we have that  $w - v^* \in L_{\text{div}}^2(\pi_0)$ . By the second item in Lemma 3, we have  $w \in \overline{\mathcal{T}_{k, \pi_0} \nabla C_c^\infty(\mathbb{R}^d)}^{\mathcal{H}_k^d}$ , and by the orthogonal decomposition (82), we see that  $\langle w - v^*, w \rangle_{\mathcal{H}_k^d} = 0$ . Rearranging, we obtain

$$\|w\|_{\mathcal{H}_k^d}^2 = \langle v^*, w \rangle_{\mathcal{H}_k^d} \leq \|v^*\|_{\mathcal{H}_k^d} \|w\|_{\mathcal{H}_k^d} \leq \|w\|_{\mathcal{H}_k^d}^2, \quad (83)$$

where we have used Cauchy-Schwarz in the first inequality and the result from Step 3 in the second inequality. The chain of inequalities in (83) forces the Cauchy-Schwarz inequality to become an equality, which implies that  $w$  and  $v^*$  are linearly dependent. Together with  $\langle w - v^*, w \rangle_{\mathcal{H}_k^d} = 0$ , it follows that  $v^* = w$ .

*Step 5.* By the previous step, any convergent sequence in the set (78) with  $t_n \rightarrow 0$  has the same (weak) limit. Therefore, the limit  $\lim_{t \rightarrow 0} \frac{1}{t} \int_0^1 v_\tau^{t_n} d\tau$  exists and equals  $v^*$ , and the proof is completed by repeating the calculation in (79).  $\square$

## D Fisher-Rao and natural gradient flows

The purpose of this appendix is to briefly explain the interpretation (35) of (34). The discussion here is purely heuristic; we refer the reader to [Ambrosio et al. \[2008\]](#), [Mielke \[2023\]](#) for rigorous accounts of gradient flow theory.

As alluded to in Section 6.2, both the gradient and the Hessian operator in (35) are interpreted in the ‘Euclidean’ way. This means that the tangent spaces

$$T_\rho \mathcal{P}(\mathbb{R}^d) = \left\{ \sigma : \int_{\mathbb{R}^d} \sigma \, dx = 0 \right\},$$

containing infinitesimal changes to  $\rho \in \mathcal{P}(\mathbb{R}^d)$  of the form  $\rho \mapsto \rho + \sigma$ , are equipped with the inner product

$$\langle \sigma_1, \sigma_2 \rangle := \int_{\mathbb{R}^d} \sigma_1 \sigma_2 \, dx, \quad (84)$$

which is similar to the standard Euclidean inner product  $\langle x, y \rangle_{\mathbb{R}^N} = \sum_{i=1}^N x_i y_i$ . The  $L^2$ -derivative of  $\rho \mapsto \text{KL}(\rho|\pi)$  at a fixed probability measure  $\rho^* \in \mathcal{P}(\mathbb{R}^d)$  can be determined from the relation

$$\partial_t \text{KL}(\rho_t|\pi) \Big|_{t=0} = \int_{\mathbb{R}^d} \nabla \text{KL}(\rho^*|\pi) \partial_t \rho_t|_{t=0} \, dx, \quad (85)$$

which is supposed to hold for all (differentiable) curves  $(\rho_t)_{t \in (-\varepsilon, \varepsilon)} \subset \mathcal{P}(\mathbb{R}^d)$  with  $\rho_0 = \rho^*$ . Here, we interpret  $\nabla \text{KL}(\rho^*|\pi)$  as an element of the dual (cotangent) space  $T_{\rho^*}^* \mathcal{P}(\mathbb{R}^d)$ , acting on the tangent vector  $\partial_t \rho_t|_{t=0} \in T_{\rho^*} \mathcal{P}(\mathbb{R}^d)$  via the  $L^2$ -pairing in (85). Direct calculation shows that

$$\partial_t \text{KL}(\rho_t|\pi) \Big|_{t=0} = \partial_t \int_{\mathbb{R}^d} \log \left( \frac{d\rho_t}{d\pi} \right) d\rho_t \Big|_{t=0} = \int_{\mathbb{R}^d} \log \left( \frac{d\rho^*}{d\pi} \right) \partial_t \rho_t|_{t=0} \, dx, \quad (86)$$

so that (85) implies

$$\nabla \text{KL}(\rho^*|\pi) = \log \left( \frac{d\rho^*}{d\pi} \right). \quad (87)$$

**Remark 12.** The derivative in (87) plays the same role as the exterior derivative [[Lee and Lee, 2012](#), Chapter 11] in differential geometry. Notice that (87) is only defined up to an additive constant, since  $\nabla \text{KL}(\rho^*|\pi)$  acts via integrals against mean-zero functions in  $T_{\rho^*} \mathcal{P}(\mathbb{R}^d)$ . We can also compute the *Riemannian gradient* with respect to (84), essentially interpreting the left-hand side of (85) as the Riemannian metric:

$$\text{grad } \text{KL}(\rho^*|\pi) = \log \left( \frac{d\rho^*}{d\pi} \right) - \int_{\mathbb{R}^d} \log \left( \frac{d\rho^*}{d\pi} \right) dx. \quad (88)$$

Importantly, (88) contains an additional constant, ensuring that  $\text{grad } \text{KL}(\rho^*|\pi) \in T_{\rho^*} \mathcal{P}(\mathbb{R}^d)$ .

In a similar manner, the Hessian (as a quadratic form on  $T_{\rho^*} \mathcal{P}(\mathbb{R}^d)$ ) can be determined from

$$\partial_t^2 \text{KL}(\rho_t|\pi)|_{t=0} = \text{Hess } \text{KL}(\rho^*|\pi)(\partial_t \rho_t|_{t=0}, \partial_t \rho_t|_{t=0}) \quad (89)$$

for all *geodesics*  $(\rho_t)_{t \in (-\varepsilon, \varepsilon)} \subset \mathcal{P}(\mathbb{R}^d)$  that satisfy  $\rho_0 = \rho^*$ . Analogously to the Euclidean case, where geodesics are given by linear interpolations of the form  $x_t = (1-t)x_0 + tx_1$ , geodesics for (84) are given by  $\rho_t = (1-t)\rho_0 + t\rho_1$ ; in particular,  $\partial_t^2 \rho_t = 0$ . As in (86), we compute

$$\partial_t^2 \text{KL}(\rho_t|\pi)|_{t=0} = \int_{\mathbb{R}^d} \frac{(\partial_t \rho)^2|_{t=0}}{\rho^*} dx + \int_{\mathbb{R}^d} \log \left( \frac{\rho_t}{\pi} \right) \underbrace{\partial_t^2 \rho_t}_{=0} dx \Big|_{t=0}.$$

Comparing to (89), we see that

$$\text{Hess } \text{KL}(\rho^*|\pi)(\sigma, \sigma) = \int_{\mathbb{R}^d} \frac{\sigma^2}{\rho^*} dx. \quad (91)$$

**Remark 13** (Fisher-Rao geometry). The Hessian quadratic form (91) coincides with the Fisher-Rao metric tensor [[Amari, 2016](#), [Ay et al., 2017](#)].

The quadratic form (91) can also be interpreted as a mapping from the tangent space  $T_{\rho^*} \mathcal{P}(\mathbb{R}^d)$  into its dual (cotangent) space  $T_{\rho^*}^* \mathcal{P}(\mathbb{R}^d)$ ,

$$T_{\rho^*} \mathcal{P}(\mathbb{R}^d) \ni \sigma \mapsto \frac{\sigma}{\rho^*} \in T_{\rho^*}^* \mathcal{P}(\mathbb{R}^d), \quad (92)$$

where we recall that the action of  $T_{\rho^*}^* \mathcal{P}(\mathbb{R}^d)$  on  $T_{\rho^*} \mathcal{P}(\mathbb{R}^d)$  is understood via the  $L^2$ -pairing in (84).<sup>7</sup> To interpret (35), we need to invert (92) and apply the result to (87). Indeed, it is immediate to verify that the required inverse<sup>8</sup> is given by

$$T_{\rho^*}^* \mathcal{P}(\mathbb{R}^d) \ni f \mapsto \rho^* \left( f - \int_{\mathbb{R}^d} f \, d\rho^* \right) \in T_{\rho^*} \mathcal{P}(\mathbb{R}^d), \quad (93)$$

so that replacing  $f$  by  $\log \left( \frac{d\rho^*}{d\pi} \right)$  connects (34) and (35).

---

<sup>7</sup>Notice that, in contrast to the situation in conventional Riemannian geometry, it is not straightforwardly possible to give meaning to the Hessian (or the Fisher-Rao metric) as an operator mapping  $T_{\rho^*} \mathcal{P}(\mathbb{R}^d)$  to itself, because  $\frac{\sigma}{\rho^*}$  does not necessarily integrate to zero.

<sup>8</sup>We highlight that (93) inverts (92) up to additive constants; functions that differ by such constants are identified in  $T_{\rho^*} \mathcal{P}(\mathbb{R}^d)$ .