

Stein Variational Gradient Descent: many-particle and long-time asymptotics

Nikolas Nüsken¹ and D.R. Michiel Renger²

¹*Institute of Mathematics, Universität Potsdam, 14476 Potsdam, Germany, nuesken@uni-potsdam.de*

²*WIAS Berlin, Mohrenstrasse 39, 10117 Berlin, Germany, renger@wias-berlin.de*

February 26, 2021

Abstract

Stein variational gradient descent (SVGD) refers to a class of methods for Bayesian inference based on interacting particle systems. In this paper, we consider the originally proposed deterministic dynamics as well as a stochastic variant, each of which represent one of the two main paradigms in Bayesian computational statistics: *variational inference* and *Markov chain Monte Carlo*. As it turns out, these are tightly linked through a correspondence between gradient flow structures and large-deviation principles rooted in statistical physics. To expose this relationship, we develop the cotangent space construction for the Stein geometry, prove its basic properties, and determine the large-deviation functional governing the many-particle limit for the empirical measure. Moreover, we identify the *Stein-Fisher information* (or *kernelised Stein discrepancy*) as its leading order contribution in the long-time and many-particle regime in the sense of Γ -convergence, shedding some light on the finite-particle properties of SVGD. Finally, we establish a comparison principle between the Stein-Fisher information and RKHS-norms that might be of independent interest.

Keywords: Stein variational gradient descent, gradient flows, large deviations.

1 Introduction

Approximating high-dimensional probability distributions is a key challenge in many applications such as Bayesian inference or computational statistical physics. The *target measure* of interest is typically given in the form

$$\pi = \frac{1}{Z} e^{-V} dx \quad (1)$$

in a high dimensional state space \mathbb{R}^d , where $Z = \int_{\mathbb{R}^d} e^{-V} dx$ is a numerically intractable normalisation constant, and $V \in C^1(\mathbb{R}^d; \mathbb{R})$ is referred to as the *potential*. Common algorithmic approaches can broadly be classified according to the following two paradigms:

Variational inference (VI) [7, 8, 81] relies on a (parameterised) family of distributions $\mathcal{D} = \{\rho_\phi : \phi \in \Phi\}$, attempting to find an approximation $\rho^* \approx \pi$ by minimising the Kullback-Leibler divergence towards the target:

$$\rho^* = \arg \min_{\rho \in \mathcal{D}} \text{KL}(\rho|\pi). \quad (2)$$

While the accuracy of VI is limited by the expressivity of \mathcal{D} , the optimisation problem (2) can often be solved efficiently and at scale using modern (stochastic) gradient descent type algorithms [33, Chapter 8].

Markov Chain Monte Carlo (MCMC) [10, 68] techniques, on the other hand, are asymptotically exact, being based on judiciously designed ergodic Markov processes $(X_t)_{t \geq 0}$ that admit π as their invariant measure. The target is obtained as an appropriate limit of a long-time ergodic average:

$$\pi = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \delta_{X_t} dt. \quad (3)$$

Accompanying convergence guarantees typically make inferences resting on MCMC more reliable than those based on VI. However, MCMC is challenging to parallelise and, furthermore, in high-dimensional settings it is often frustrated by slow convergence in (3) due to time correlations in $(X_t)_{t \geq 0}$.

Recently, there has been a growing interest in developing hybrid approaches that hold the promise of combining the advantages of MCMC and VI, see, for instance [36, 51, 55, 70, 72]. Various attempts in this direction can be grouped into the so-called *particle optimisation techniques* [1, 11, 12, 43] that posit carefully designed dynamical schemes for an ensemble of particles $\bar{X} = (X^1, \dots, X^N) \in (\mathbb{R}^d)^N$. From the VI-perspective, the variational family is then given by the empirical measures associated to the particles, $\mathcal{D} = \{\frac{1}{N} \sum_{i=1}^N \delta_{X^i}\}$, with the parameter set Φ corresponding to the positions of these particles. In terms of MCMC, the dynamics of $(\bar{X}_t)_{t \geq 0}$ can often be at least approximately thought of as a Markov process approaching an extended target $\bar{\pi}$ on $(\mathbb{R}^d)^N$ whose marginals coincide with π .

An appealing theoretical framework for analysing and constructing these particle-based methods is provided by the theory of gradient flows on probability distributions [2, 60, 62], connecting diffusions with KL-optimisation problems of the form (2) on the grounds of differential geometric ideas. In this regard, the prime example (and also historically the first one where these concepts were laid out, see [37]) is given by the overdamped Langevin dynamics [64, Section 4.5], the associated Fokker-Planck equation of which takes the form of a gradient flow evolution driven by the KL-divergence in the geometry induced by the quadratic Wasserstein distance. Recently, similar ideas have been pursued, replacing either the driving functional or the underlying geometry, see, for instance, [3, 23, 29, 30, 44, 67, 77].

In statistical physics, gradient flow structures have been shown to play a major role in understanding the fluctuations of associated (stochastic) interacting particle systems [54, 57, 58] as described by the theory of large deviations. In this paper we utilise the correspondence between gradient flow structures and large-deviation functionals to shed some light on the connection between VI and MCMC in the context of a particular particle optimisation scheme, namely Stein variational gradient descent.

1.1 Stein Variational Gradient Descent

Following the VI-paradigm, *Stein variational gradient descent* (SVGD) was first derived in [46] from a minimising movement scheme for an ensemble of particles, seeking to iteratively solve the problem (2) for the corresponding empirical measure, while at the same time constraining the driving vector field to be chosen from within the unit ball of a reproducing kernel Hilbert space (RKHS)¹. The method can be described by the following coupled system of ODEs, where $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a positive definite kernel of sufficient regularity² and $\bar{X}_t = (X_t^1, \dots, X_t^N) \in (\mathbb{R}^d)^N$ denotes the ensemble of particles:

$$\frac{dX_t^i}{dt} = \frac{1}{N} \sum_{j=1}^N \left(-k(X_t^i, X_t^j) \nabla V(X_t^j) + \nabla_{X_t^j} k(X_t^i, X_t^j) \right), \quad i = 1, \dots, N. \quad (4)$$

Crucial to this approach is the observation that the corresponding empirical measure

$$\rho_t^{(N)} := \frac{1}{N} \sum_{i=1}^N \delta_{X_t^i} \quad (5)$$

converges to the target π in an appropriate sense as both $N \rightarrow \infty$ and $t \rightarrow \infty$, see [49] for rigorous statements. In [28], the authors proposed to augment (4) and obtained the interacting system of stochastic differential equations (SDEs)

$$dX_t^i = \frac{1}{N} \sum_{j=1}^N \left[-k(X_t^i, X_t^j) \nabla V(X_t^j) + \nabla_{X_t^j} k(X_t^i, X_t^j) \right] dt + \sum_{j=1}^N \sqrt{2\mathcal{K}(\bar{X}_t)_{ij}} dW_t^j, \quad i = 1, \dots, N, \quad (6)$$

where the matrix-valued function $\mathcal{K} : (\mathbb{R}^d)^N \rightarrow \mathbb{R}^{dN \times dN}$ consists of N^2 blocks of size $d \times d$, given by $\mathcal{K}_{ij}(\bar{x}) = \frac{1}{N} k(x_i, x_j) I_{d \times d}$, for $i, j \in \{1, \dots, d\}$ and $\bar{x} = (x_1, \dots, x_N)$. Here, $(W_t^j)_{t \geq 0}$, $j = 1, \dots, N$ denotes a collection of

¹Even though the KL-divergence between the empirical measure and π is not defined (or infinite), this statement can be made precise using the closely related *kernelised Stein discrepancy* [45].

²We refer to Section 2 for precise assumptions.

d -dimensional standard Brownian motions, and $\sqrt{\mathcal{K}(\bar{X}_t)}$ refers to the matrix square root. The noise contribution $\sum_{j=1}^N \sqrt{2\mathcal{K}(\bar{X}_t)_{ij}} dW_t^j$ has been designed so as to make the product measure $\bar{\pi}$ on $(\mathbb{R}^d)^N$ with Lebesgue density

$$\bar{\pi}(\bar{x}) = \pi(x_1) \cdot \dots \cdot \pi(x_N), \quad \bar{x} = (x_1, \dots, x_N), \quad (7)$$

invariant for the dynamics (6). In fact, under reasonable assumptions, [23, Proposition 3] shows that (6) is indeed ergodic with respect to $\bar{\pi}$, meaning that the associated empirical measure converges to π as $t \rightarrow \infty$ (for instance, in total variation distance). These observations show that the process $(\bar{X}_t)_{t \geq 0}$ solving (6) can indeed be considered of MCMC-type, targeting $\bar{\pi}$. Indeed, (6) can be cast in the framework of [50] as pointed out in [28].

1.2 A connection between MCMC and VI rooted in statistical physics

One of the main topics in this article is the connection between the ODE (4) and the SDE (6). First of all, the empirical measures associated to the solutions of (4) and (6) become indistinguishable in the limit as $N \rightarrow \infty$, that is, the noise term $\sum_{j=1}^N \sqrt{2\mathcal{K}(\bar{X}_t)_{ij}} dW_t^j$ becomes negligible. This claim can be substantiated in the sense that the Stein PDE [44, 49]

$$\partial_t \rho_t(x) = \nabla_x \cdot \left(\rho_t(x) \int_{\mathbb{R}^d} [k(x, y) \nabla V(y) - \nabla_y k(x, y)] \rho_t(dy) \right) \quad (8)$$

describes the evolution of the empirical measure $\rho_t^{(N)}$ for both (4) and (6) in the mean field regime, that is, when $N \rightarrow \infty$. To be more precise, for any fixed $N \in \mathbb{N}$, the empirical measure $\rho_t^{(N)}$ associated to the ODE (4) satisfies (8) in a weak sense, see [49, Prop. 2.5], and, moreover, stability arguments show that this statement can be extended to the limit $N \rightarrow \infty$, see [49, Theorem 2.7]. Concerning the SDE (6), the additional noise term has been shown to be of order $\mathcal{O}(\frac{1}{N})$ in [23, Proposition 3] and thus the corresponding empirical measure $\rho_t^{(N)}$ formally satisfies (8) in the limit as $N \rightarrow \infty$. In this paper the latter convergence will be made more quantitative in terms of a corresponding large-deviation functional.

The Stein PDE (8) admits a gradient flow structure, described in [44] and further analysed in [23], that is, it can be written in the form $\partial_t \rho_t = -\text{grad}_k \text{KL}(\rho_t)$, where KL is the Kullback-Leibner divergence or relative entropy towards π , and the gradient is with respect to a particular geometry determined by the kernel k ; we shall make these terms more precise in Section 3. The gradient flow structure referred to above is not uniquely determined by (8); in fact the existence of one particular gradient flow structure implies that the PDE (8) admits infinitely many gradient flow structures [20]. For a particular example see [13], replacing the KL- by the χ^2 -divergence. Our first main result shows that the KL-gradient flow structure is naturally connected to the noise contribution in (6), bridging between the MCMC and VI viewpoints:

Informal Result 1.1. *The gradient flow structure*

$$\partial_t \rho_t = -\text{grad}_k \text{KL}(\rho_t) \quad (9)$$

for the Stein PDE (8) (see Section (3)) is compatible with the particular form of the noise in the SDE (6).

This statement will be made precise in Section 5, resting on a reformulation of the Stein PDE (8) in terms of a variational (in-)equality (see Proposition 3.11) and the large-deviation functional for the $N \rightarrow \infty$ limit associated to the SDE (6), see Theorem 4.3. Intuitively, both the gradient flow scheme (9) and the large-deviation functional related to the noise structure in (6) encode information that goes beyond what is described by the Stein PDE (8): The formulation (9) determines a specific non-unique ‘factorisation’ of the right-hand side of (8) into the geometric term grad_k and the driving functional KL, while the SDE (6) determines a non-unique³ stochastic augmentation of (4). The Informal Result 1.1 establishes a correspondence between those extensions of (8) rooted in statistical physics; this general principle can be seen as a modern version of Onsager’s reciprocity relation [53, 57].

1.3 Speed of convergence, kernel choice and Stein-Fisher information

From the practical perspective of minimising the computational cost, a central question is how to choose k in such a way that the convergence $\rho_t^{(N)} \rightarrow \pi$ as $N \rightarrow \infty$ and $t \rightarrow \infty$ occurs ‘as rapidly as possible’, that is, in such a way that $\rho_t^{(N)}$ can provide a reasonable approximation of π for t and N not too large. In [23], the authors used convexity

³However, the noise contribution in (6) is canonical in Bayesian inference as it ensures ergodicity with respect to the extended target $\bar{\pi}$.

arguments along the geodesics induced by the Stein geometry for studying the $t \rightarrow \infty$ limit of the PDE (8), that is, for the study of the long-time behaviour in the many-particle regime.

In the present paper, we complement those results, quantifying the speed of convergence for the random dynamics (6) as $N \rightarrow \infty$ using the theory of large deviations (see Section 4). As a consequence of the Informal Result 1.1, the relevant functional admits an elegant formulation in terms of the Stein geometry (see Theorem 4.3). Although the methods presented in this paper concern the SDE system (6), these results allow us to gain some intuition into finite-particle effects for the deterministic system (4) on a heuristic level (see Section 6.2).

In order to gain further insight and in particular to derive practical guidelines for the choice of k , we next identify the leading order term in the large-deviation functional when t is large, where limits are performed in the sense of Gamma convergence. In this regime, both the many-particle as well as the long-time asymptotics turn out to be closely related to the *Stein-Fisher information*

$$I_{\text{Stein}}^k(\rho) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \nabla \frac{d\rho}{d\pi}(x) \cdot k(x, y) \nabla \frac{d\rho}{d\pi}(y) \pi(dx) \pi(dy) \quad (10a)$$

$$= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \left(\nabla \log \frac{d\rho}{d\pi} \right)(x) \cdot k(x, y) \left(\nabla \log \frac{d\rho}{d\pi} \right)(y) \pi(dx) \pi(dy), \quad (10b)$$

a quantity that has natural links with the cotangent space construction to be introduced in Section 3.2. Let us also note that $I_{\text{Stein}}^k(\rho)$ is known in other contexts as the *kernelised Stein discrepancy* $\text{KSD}(\rho|\pi)$ and has found various applications in scenarios where ρ needs to be compared to an unnormalised⁴ distribution π , see [14, 25, 34]. In fact, the kernelised Stein discrepancy lies at the heart of the original derivation of SVGD, see [46]. We summarise our findings in the following informal statement (to be explained and justified in Section 6).

Informal Result 1.2. *The Stein-Fisher information I_{Stein}^k controls the speed of convergence of the empirical measure associated to the SDE (6) in the regime when N and t are large. As a consequence, letting $k_1, k_2 : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be two positive definite kernels with corresponding empirical measures $\rho_t^{(N), k_1}$ and $\rho_t^{(N), k_2}$ as defined in (5) and (6), if*

$$I_{\text{Stein}}^{k_1}(\rho) \geq I_{\text{Stein}}^{k_2}(\rho), \quad (11)$$

for all ρ such that (10a) is well defined, then the convergence of $\rho_t^{(N), k_1}$ towards π as $N \rightarrow \infty$ and $t \rightarrow \infty$ is expected to be faster than the corresponding convergence of $\rho_t^{(N), k_2}$.

The preceding result applies when N is large, but not infinite, hence taking a step towards understanding the finite-particle properties of SVGD. Naturally, our two main results are strongly related. Indeed, the fact that the Stein-Fisher information (10) controls the speed of convergence in both the $t \rightarrow \infty$ and $N \rightarrow \infty$ limits is ultimately a consequence of the compatibility between the gradient flow and noise structures expressed in the Informal Result 1.1. Let us state straight away that the comparison (11) can be made on the basis of the reproducing kernel Hilbert spaces (RKHS) associated to k_1 and k_2 . More precisely, we shall prove the following result (see Section 6).

Proposition 1.3. *Let $k_1, k_2 : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be two positive definite kernels satisfying Assumptions 2, 3, and 4 below, and denote by \mathcal{H}_{k_1} and \mathcal{H}_{k_2} the corresponding reproducing kernel Hilbert spaces. Furthermore, assume that V satisfies Assumption 1. Then the following are equivalent:*

1. Stein-Fisher comparison: *The inequality (11) holds for all ρ such that (10a) is well defined,*
2. Inclusion of RKHS-balls: *It holds that $\mathcal{H}_{k_2} \subset \mathcal{H}_{k_1}$ and*

$$\|\phi\|_{\mathcal{H}_{k_2}} \leq \|\phi\|_{\mathcal{H}_{k_1}}, \quad \text{for all } \phi \in \mathcal{H}_{k_1}. \quad (12)$$

We refer the reader to Section 6.2 for a proof, and to Section 7 for an illustration of this result. Noting that the Stein-Fisher information coincides with the kernelised Stein discrepancy $\text{KSD}(\rho|\pi)$, Proposition 1.3 might be of independent interest.

⁴Indeed, (10b) shows that $I_{\text{Stein}}^k(\rho)$ can be computed from $\pi = \frac{1}{Z} e^{-V}$ without knowing the potentially intractable normalisation constant Z .

Previous work

Stein variational gradient descent in its original deterministic form (4) was put forward in the seminal paper [46]. The stochastic variant (6) was proposed in [28] and shown to be ergodic in [23]. The fact that the Stein PDE (8) admits a gradient flow structure was first observed in [44]; the corresponding Stein geometry was further developed in [23], focusing on curvature and the long-time convergence properties of (8). This analysis revealed the important role played by the Stein-Fisher information (10) and the associated Stein log-Sobolev inequality. Based on this, the authors of [39] developed nonasymptotic bounds in discrete time as well as propagation of chaos results (the latter of which unfortunately are not uniform in time). We would also like to mention the work [49] that rigorously establishes well-posedness as well as convergence of the Stein PDE (8), and the work [13] that establishes an alternative gradient flow structure to the one considered in this paper.

Our contributions and outline of the article

In this article we make the following contributions:

- We complement the geometric constructions from [23], defining appropriate cotangent spaces and inner products. The Stein-Fisher information (10) (or kernelised Stein discrepancy) is shown to have a natural interpretation in terms of this framework.
- We compute the large-deviation functional associated to the mean field limit of the SDE (6) and show that it can be expressed conveniently in terms of the tangent norm in the Stein geometry.
- On the basis of the obtained large-deviation rate functional, we connect the KL-gradient flow structure in (9) with the noise structure in (6), providing a correspondence between the VI-type scheme (4) and the MCMC-type scheme (6).
- We identify the leading order term in the large-deviation functional in the regime where t is large to obtain a direct relation to the Stein-Fisher information (10). We argue that at a heuristic level, this result provides insight into finite-particle properties of SVGD.

The article is organised as follows. In Section 2 we introduce essential notation and state our basic assumptions. Furthermore, we provide an overview of the relevant background on reproducing kernel Hilbert spaces. In Section 3.1, we review the geometric constructions from [23]. In Section 3.2, we extend this work by defining the cotangent structure and establish its basic properties. Furthermore, we provide a reformulation of the Stein PDE (8) in terms of a variational (in-)equality. In Section 4 we derive the large-deviation rate functional for the mean field limit, leveraging the framework introduced in Section 3. In Section 5, we explain the connection between gradient flows and large deviations and make the Informal Result 1.1 precise. In Section 6, we identify the Stein-Fisher information as the leading order term in the large-deviation rate functional, provide a precise statement of the Informal Result 1.2, and prove Proposition 1.3. Furthermore, we provide a numerical example that illustrates our results. Finally, we conclude the paper in Section 7 and briefly discuss directions for future work.

2 Preliminaries

In this section, we introduce essential notations and assumptions that are used throughout this article. In addition, we briefly point out a few key results in the theory of reproducing kernel Hilbert spaces. For textbook accounts, the reader is referred to [4, 71, 73, 76].

2.1 Notation and general assumptions

In order to ensure that both the target measure π in (1) as well as the dynamics (4) and (6) are well-defined, we assume that the given potential satisfies the following:

Assumption 1 (Assumptions on V). *The potential V is continuously differentiable, $V \in C^1(\mathbb{R}^d)$, and e^{-V} is integrable, $\int_{\mathbb{R}^d} e^{-V} dx < \infty$.*

The set of probability measures on \mathbb{R}^d will be denoted by $\mathcal{P}(\mathbb{R}^d)$. For any $\rho \in \mathcal{P}(\mathbb{R}^d)$, the Hilbert space of ρ -square-integrable functions will be denoted by $L^2(\rho)$, with scalar product $\langle \phi, \psi \rangle_{L^2(\rho)} = \int_{\mathbb{R}^d} \phi \psi d\rho$ and associated norm

$\|\phi\|_{L^2(\rho)}^2 = \langle \phi, \phi \rangle_{L^2(\rho)}$. Often, we will work with the following subset of probability measures,

$$M := \left\{ \rho \in \mathcal{P}(\mathbb{R}^d) : \rho \text{ admits a smooth and strictly positive density with respect to the Lebesgue measure} \right\}. \quad (13)$$

We later formally turn this set into a Riemannian manifold with an extended geodesic distance (allowing the value ∞) that depends on the choice of the kernel.

2.2 Assumptions on kernels

Throughout this paper, we work with one or more kernels that are always assumed to satisfy the following:

Assumption 2. *The kernel $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is assumed to be symmetric, continuous, and continuously differentiable off the diagonal, that is, $k \in C^1(\mathbb{R}^d \times \mathbb{R}^d \setminus \{(x, y) \in \mathbb{R}^{2d} : x = y\})$. Furthermore, k is assumed to be positive definite, that is, for all $n \in \mathbb{N}$, $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ and $x_1, \dots, x_n \in \mathbb{R}^d$ it holds that $\sum_{i,j \geq 1}^n \alpha_i \alpha_j k(x_i, x_j) \geq 0$.*

Assumption 3. *The kernel k is bounded.*

Assumption 4. [27, 75] *The kernel k is integrally strictly positive definite (ISPD), that is,*

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} k(x, y) \rho(dx) \rho(dy) > 0, \quad (14)$$

for all signed Borel measures ρ that are not the zero measure.

Let us comment on the foregoing assumptions. While Assumption 2 is fundamental (in that it is required for the construction of associated reproducing kernel Hilbert spaces (RKHS) as well as for defining all the terms in (4) and (6)), Assumptions 3 and 4 are made in this paper for technical convenience. Indeed, the set-up in [23] encompasses unbounded kernels (but does require the weaker integrability condition $\int_{\mathbb{R}^d} k(x, x) d\rho(x) < \infty$ for measures ρ under consideration). Non-ISPD kernels have been considered in [47], for instance, and could be included in our framework with more technical effort. Note that the ISPD Assumption 4 is a strengthened version of the positive definiteness in Assumption 2.

Examples of kernels satisfying Assumptions 2, 3 and 4 are given by the parametric family $k_{p,\sigma} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, defined via

$$k_{p,\sigma}(x, y) = \exp\left(-\frac{|x - y|^p}{\sigma^p}\right), \quad (15)$$

where $p \in (0, 2]$ is a smoothness parameter, and $\sigma > 0$ is called the kernel width (see [23, Lemma 42]). Further examples are provided by the family of *Matérn kernels* whose reproducing kernel Hilbert spaces coincide with the classical Sobolev spaces $W^{m,2}(\mathbb{R}^d)$ whenever m and d are such that $W^{m,2}(\mathbb{R}^d) \subset C(\mathbb{R}^d)$, see [71, Section 1.3].

2.3 Reproducing kernel Hilbert spaces

Given a positive definite kernel k , we denote by $(\mathcal{H}_k, \langle \cdot, \cdot \rangle_{\mathcal{H}_k})$ the corresponding reproducing kernel Hilbert space (RKHS), see [76, Section 4], and by $\|\cdot\|_{\mathcal{H}_k} = \langle \cdot, \cdot \rangle_{\mathcal{H}_k}^{1/2}$ the associated norm. This Hilbert space is characterised by the conditions that $k(x, \cdot) \in \mathcal{H}_k$ as well as $\langle f, k(x, \cdot) \rangle_{\mathcal{H}_k} = f(x)$, for all $x \in \mathbb{R}^d$ and $f \in \mathcal{H}_k$. If $\rho \in \mathcal{P}(\mathbb{R}^d)$ is a probability measure with full support, then Assumption 3 ensures that $\mathcal{H}_k \subset L^2(\rho)$, where moreover the natural inclusion is continuous, see [76, Theorem 4.26], and Assumption 4 guarantees that $\mathcal{H}_k \subset L^2(\rho)$ is dense, see [75, Theorem 7] and [76, Theorem 4.26i)].

In order to characterise the norm $\|\cdot\|_{\mathcal{H}_k}$ more explicitly, it is helpful to introduce the operators $\mathcal{T}_{k,\rho} : L^2(\rho) \rightarrow L^2(\rho)$

$$(\mathcal{T}_{k,\rho}\phi)(x) := \int_{\mathbb{R}^d} k(x, y)\phi(y)\rho(dy), \quad \phi \in L^2(\rho). \quad (16)$$

We gather a number of properties of this operator that will be useful later on.

Proposition 2.1. *For all $\rho \in M$,*

(a) $\mathcal{T}_{k,\rho} L^2(\rho) \subset \mathcal{H}_k$, and $\mathcal{T}_{k,\rho} : L^2(\rho) \rightarrow \mathcal{H}_k$ is the adjoint of the inclusion $\mathcal{H}_k \hookrightarrow L^2(\rho)$, that is

$$\langle \mathcal{T}_{k,\rho}\phi, \psi \rangle_{\mathcal{H}_k} = \langle \phi, \psi \rangle_{L^2(\rho)}, \quad \phi \in L^2(\rho), \psi \in \mathcal{H}_k. \quad (17)$$

(b) $\mathcal{T}_{k,\rho}$ is compact, self-adjoint and positive semi-definite on $L^2(\rho)$,

(c) $\mathcal{T}_{k,\rho}$ is injective.

Proof. For (a) and (b), see [76], Theorems 4.26 and 4.27, respectively. For (c), notice that $\mathcal{T}_{k,\rho}\phi = 0$, $\phi \in C_c^\infty(\mathbb{R}^d)$ implies $\phi = 0$ by integrating against $\phi\rho$ and using Assumption 4. \square

Remark 2.2. The identity (17) is a key calculational tool throughout the proofs in Section 3 and can formally be viewed as a consequence of the defining identity $\langle k(x, \cdot), f \rangle_{\mathcal{H}_k} = f(x)$ after commuting integration and $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$.

The scalar product in \mathcal{H}_k can now be written in the form

$$\langle f, g \rangle_{\mathcal{H}_k} = \langle \mathcal{T}_{k,\rho}^{-1/2} f, \mathcal{T}_{k,\rho}^{-1/2} g \rangle_{L^2(\rho)}, \quad f, g \in \mathcal{H}_k, \quad (18)$$

where $\mathcal{T}_{k,\rho}^{-1/2}$ may be defined via the spectral theorem [65, Chapter VII]. For instance, if $(e_i)_{i=1}^\infty \subset L^2(\rho)$ is an orthonormal eigenbasis of $\mathcal{T}_{k,\rho}$ (that is, $\langle e_i, e_j \rangle_{L^2(\rho)} = \delta_{ij}$ and $\mathcal{T}_{k,\rho} e_i = \lambda_i e_i$), then for $f = \sum_i f_i e_i$ and $g = \sum_i g_i e_i$ we have that

$$\langle f, g \rangle_{\mathcal{H}_k} = \sum_{i=1}^{\infty} \frac{1}{\lambda_i} f_i g_i,$$

see [76, Section 4.5].

Derived from \mathcal{H}_k and $L^2(\rho)$, we will frequently make use of the corresponding spaces of vector fields \mathcal{H}_k^d and $(L^2(\rho))^d$, defined through

$$\mathcal{H}_k^d = \underbrace{\mathcal{H}_k \otimes \dots \otimes \mathcal{H}_k}_{d \text{ times}} \quad \text{and} \quad (L^2(\rho))^d = \underbrace{L^2(\rho) \otimes \dots \otimes L^2(\rho)}_{d \text{ times}}.$$

In other words, \mathcal{H}_k^d and $(L^2(\rho))^d$ consist of vector fields $v = (v_1, \dots, v_d)$, with $v_i \in \mathcal{H}_k$ or $v_i \in L^2(\rho)$, respectively, with scalar products given by

$$\langle v, w \rangle_{\mathcal{H}_k^d} = \sum_{i=1}^d \langle v_i, w_i \rangle_{\mathcal{H}_k}, \quad v_i, w_i \in \mathcal{H}_k, \quad \text{and} \quad \langle v, w \rangle_{(L^2(\rho))^d} = \sum_{i=1}^d \langle v_i, w_i \rangle_{L^2(\rho)}, \quad v_i, w_i \in L^2(\rho).$$

The operators $\mathcal{T}_{k,\rho}$ defined in (16) straightforwardly extend to the space $(L^2(\rho))^d$, interpreting (16) componentwise. Similarly, Proposition 2.1 as well as the identity (18) remain valid with the obvious modifications. Finally, we will need the following result in the spirit of the usual Helmholtz-decomposition [74].

Proposition 2.3 (Helmholtz decomposition for RKHS). *Let $\rho \in M$ and define the space of divergence-free vector fields*

$$L_{\text{div}}^2(\rho) = \{v \in (L^2(\rho))^d : \langle v, \nabla \phi \rangle_{(L^2(\rho))^d} = 0, \quad \text{for all } \phi \in C_c^\infty(\mathbb{R}^d)\}. \quad (19)$$

Then \mathcal{H}_k^d admits the following $\langle \cdot, \cdot \rangle_{\mathcal{H}_k^d}$ -orthogonal decomposition,

$$\mathcal{H}_k^d = (L_{\text{div}}^2(\rho) \cap \mathcal{H}_k^d) \oplus \overline{\mathcal{T}_{k,\rho} \nabla C_c^\infty(\mathbb{R}^d)}^{\mathcal{H}_k^d}. \quad (20)$$

Proof. We refer to [23, Lemma 45]. \square

3 The Stein PDE as a gradient flow

In this section we recall and further analyse the *Stein geometry* that allows us to formally write the Stein PDE (8) as a gradient flow on probability distributions, as first observed in [44]. Subsection 3.1 will mostly be a review of the Stein geometry as developed in [23]; in Subsection 3.2 we complement the construction from [23] by defining appropriate cotangent spaces endowed with inner products; those turn out to be closely related to the Stein-Fisher information (10). The duality between tangent and cotangent spaces gives rise to a variational reformulation of the Stein PDE (8) in Proposition 3.11 that will be instrumental in linking the large-deviation statement in Section 4 to the gradient flow structure of the mean field limit. Analysing the Stein PDE (8) using the geometric picture outlined in this section is very much inspired by the works of Otto and coworkers on the Fokker-Planck equation and its relation to the quadratic Wasserstein distance (see [37, 59, 60, 61, 62] as well as the further developments in [2, 32] and [15]). For a direct comparison between the Stein geometry and the Wasserstein geometry we refer the reader to [23, Appendix A].

Anticipating the constructions to follow in the remainder of this section, let us already in intuitive terms lay out the connections between the original idea from [46] and the central geometric concepts of the Stein geometry. In [46], the authors construct the ODE (4) as the continuous-time limit of a gradient descent scheme. More precisely, they consider an ensemble of particles, represented by the empirical measure $\rho^{(N)}$, and design a minimising movement scheme that aims at minimising the KL-divergence between $\rho^{(N)}$ and the target π . The associated velocity field is constrained to be chosen from within the RKHS \mathcal{H}_k^d and obtained from a variational principle that involves the corresponding RKHS-norm. As observed in [44] and further developed in [23], this construction principle is linked to the observation that (8) can be cast in the form

$$\partial_t \rho = -\mathbb{K}_\rho \frac{\delta \text{KL}}{\delta \rho} =: -(\text{grad}_k \text{KL})(\rho), \quad (21)$$

where KL denotes the Kullback-Leibler⁵ divergence (or relative entropy) between the current distribution ρ_t and the target π ,

$$\text{KL}(\rho) = \int_{\mathbb{R}^d} \log \left(\frac{d\rho}{d\pi} \right) d\rho = \int_{\mathbb{R}^d} V d\rho + \int_{\mathbb{R}^d} \log \rho d\rho + \log Z, \quad \rho \in M, \quad (22)$$

and \mathbb{K}_ρ is a positive definite ‘Onsager’ operator that we introduce in (29b). This operator defines the Stein-gradient $\text{grad}_k := \mathbb{K}_\rho \frac{\delta}{\delta \rho}$, formalises the minimising movement scheme from [46] and can be seen to be induced by an appropriate definition of the tangent spaces $T_\rho M$ and corresponding (formal) Riemannian metric. The Onsager operators \mathbb{K}_ρ translate between the tangent and cotangent spaces defined below; indeed we have $\text{grad}_k \text{KL} \in T_\rho M$ and $\frac{\delta \text{KL}}{\delta \rho} \in T_\rho^* M$, at least formally.

From a statistical perspective, the term $\int_{\mathbb{R}^d} V d\rho$ in (22) measures the fit to the data, the entropic term $\int_{\mathbb{R}^d} \log \rho d\rho$ encodes regularisation, and the normalisation constant Z represents the Bayesian evidence, useful in the context of model selection (see, for instance [48] and [31, Section 6.7]).

3.1 Formal Riemannian structure and associated gradient

In what follows, we formally equip the set M defined in (13) with the structure of a Riemannian manifold, following [23, Section 4], where the reader is referred to for further details. To start with, recall the operators $\mathcal{T}_{k,\rho}$ from (16), that can be extended to self-adjoint, nonnegative definite, and compact operators on $L^2(\rho)$, see [76, Section 4.3]. By abuse of notation, we will often apply $\mathcal{T}_{k,\rho}$ to vector fields in $(L^2(\rho))^d$, in which case (16) is to be understood componentwise. The operators $\mathcal{T}_{k,\rho}$ are used to define the tangent space construction in the Stein geometry:

Definition 3.1 (Tangent spaces and Riemannian metric). See [23, Definition 5]. For $\rho \in M$, we define the *tangent space*⁶

$$T_\rho M := \left\{ \xi \in \mathcal{D}'(\mathbb{R}^d) : \text{there exists } v \in \overline{\mathcal{T}_{k,\rho} \nabla C_c^\infty(\mathbb{R}^d)}^{\mathcal{H}_k^d} \text{ such that } \xi + \nabla \cdot (\rho v) = 0 \text{ in the sense of distributions} \right\}$$

and the *Riemannian metric* $\langle \cdot, \cdot \rangle_{T_\rho M} : T_\rho M \times T_\rho M \rightarrow \mathbb{R}$ by

$$\langle \xi, \chi \rangle_{T_\rho M} := \langle u, v \rangle_{\mathcal{H}_k^d}, \quad (24)$$

where $\xi + \nabla \cdot (\rho u) = 0$ and $\chi + \nabla \cdot (\rho v) = 0$, as well as $u, v \in \overline{\mathcal{T}_{k,\rho} \nabla C_c^\infty(\mathbb{R}^d)}^{\mathcal{H}_k^d}$.

A few remarks concerning Definition 3.1 are in order. First of all, the spaces $\overline{\mathcal{T}_{k,\rho} \nabla C_c^\infty(\mathbb{R}^d)}^{\mathcal{H}_k^d}$ mimic the spaces $\overline{\nabla C_c^\infty(\mathbb{R}^d)}^{L^2(\rho)}$ common in the Wasserstein setting, see for example [2, Sec. 8.4]. Similar to that scenario, to each $\xi \in T_\rho M$ there exists a unique $u \in \overline{\mathcal{T}_{k,\rho} \nabla C_c^\infty(\mathbb{R}^d)}^{\mathcal{H}_k^d}$ with $\xi + \nabla \cdot (\rho u) = 0$, so that (24) is justified. This fact can be traced back to the Helmholtz decomposition in the RKHS setting, see Proposition 2.3. Furthermore, $\overline{\mathcal{T}_{k,\rho} \nabla C_c^\infty(\mathbb{R}^d)}^{\mathcal{H}_k^d}$ may also be recognised as the set of vector fields which are permissible in minimising movement

⁵For notational convenience later on, we adopt the notation $\text{KL}(\rho) := \text{KL}(\rho|\pi)$, suppressing the dependence on π .

⁶ \mathcal{D}' denotes the usual space of Schwartz distributions, as the dual of $\mathcal{D}(\mathbb{R}^d) := C_c^\infty(\mathbb{R}^d)$ equipped with the Schwartz topology, see [22]. Moreover, we say that $\xi + \nabla \cdot (\rho v) = 0$ holds in the sense of distributions if $\langle \xi, \phi \rangle - \int_{\mathbb{R}^d} \nabla \phi \cdot v d\rho = 0$ for all $\phi \in C_c^\infty(\mathbb{R}^d)$, where $\langle \cdot, \cdot \rangle$ denotes the standard duality relation between $\mathcal{D}'(\mathbb{R}^d)$ and $C_c^\infty(\mathbb{R}^d)$.

schemes such as those devised in the original paper [44]. Therefore, at an intuitive level, $T_\rho M$ is the space of derivatives $\partial_t \rho$, where ρ is a curve obtained by continuous-time limits of these schemes. We refer the reader to [23, Lemma 7], showing that indeed $T_\rho M$ is a well-defined Hilbert space, for all $\rho \in M$.

The following lemma shows that $T_\rho M$ can indeed be considered the tangent space. Before we come to this result, we recall that the functional derivative of a suitable functional $\mathcal{F} : M \rightarrow \mathbb{R}$ is defined via

$$\int_{\mathbb{R}^d} \frac{\delta \mathcal{F}}{\delta \rho}(\rho)(x) \phi(x) dx := \left. \frac{d}{d\varepsilon} \right|_{\varepsilon=0} \mathcal{F}(\rho + \varepsilon \phi), \quad (25)$$

for $\phi \in C_c^\infty(\mathbb{R}^d)$ with $\int_{\mathbb{R}^d} \phi dx = 0$. The functional derivative of the Kullback-Leibner divergence (22) can be computed for $\rho \in M$:

$$\frac{\delta \text{KL}}{\delta \rho}(\rho)(x) = \log \rho(x) + V(x), \quad (26)$$

see, for instance, [80, Chapter 15].

Finally, we are able to connect the geometric construction from Definition 3.1 with the Stein PDE (8):

Lemma 3.2 (Stein gradient). *See [23, Lemma 9 and Corollary 11]. Let $\rho \in M$ and $\mathcal{F} : M \rightarrow \mathbb{R}$ be such that the functional derivative $\frac{\delta \mathcal{F}}{\delta \rho}(\rho)$ is well-defined and continuously differentiable. Moreover assume that $\mathcal{T}_{k,\rho} \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\rho) \in \overline{\mathcal{T}_{k,\rho} \nabla C_c^\infty(\mathbb{R}^d)}^{\mathcal{H}^k}$. Then the Riemannian gradient associated to $(T_\rho M, \langle \cdot, \cdot \rangle_{T_\rho M})$ is given by*

$$(\text{grad}_k \mathcal{F})(\rho) = -\nabla \cdot \left(\rho \mathcal{T}_{k,\rho} \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\rho) \right). \quad (27)$$

Using (26), it follows that the gradient flow formulation (21) and the Stein PDE (8) coincide.

3.2 Cotangent spaces, Onsager operators, duality and the energy-dissipation (in-)equality

In this section we expand the ideas of [23] and define the cotangent spaces $T_\rho^* M$, their duality relationship with the tangent spaces through the Onsager operators \mathbb{K}_ρ , and establish their basic properties. In order to construct the cotangent spaces, we begin by defining the corresponding inner products for sufficiently regular test functions.

Definition 3.3 (Dual inner product). For $\rho \in M$, we define the *dual inner product*

$$\langle \phi, \psi \rangle_{T_\rho^* M} = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \nabla \phi(x) \cdot k(x, y) \nabla \psi(y) \rho(dx) \rho(dy), \quad \phi, \psi \in C_c^\infty(\mathbb{R}^d), \quad (28)$$

as well as the *Onsager operator*

$$\mathbb{K}_\rho : C_c^\infty(\mathbb{R}^d) \rightarrow T_\rho M \quad (29a)$$

$$\phi \mapsto -\nabla \cdot (\rho \mathcal{T}_{k,\rho} \nabla \phi). \quad (29b)$$

Remark 3.4. Combining the definition (29) with (27), we see that $\text{grad}_k = \mathbb{K}_\rho \frac{\delta}{\delta \rho}$. In differential geometric terms, the functional derivative $\frac{\delta}{\delta \rho}$ takes the role of the exterior derivative [41], while the Onsager operator corresponds to the musical isomorphisms (‘raising’ the index in the language of theoretical physics). The latter concept will be made more explicit in Proposition 3.8 below. Note also that (29b) is similar to the Wasserstein setting where $\mathbb{K}_\rho \phi = -\nabla \cdot (\rho \nabla \phi)$.

The next lemma is a prelude to Definition 3.6, in particular showing that the inner product $\langle \cdot, \cdot \rangle_{T_\rho^* M}$ is nondegenerate.

Lemma 3.5. *Let k satisfy Assumptions 2, 3 and 4. Then $(C_c^\infty(\mathbb{R}^d), \langle \cdot, \cdot \rangle_{T_\rho^* M})$ is a pre-Hilbert⁷ space over \mathbb{R} .*

Proof. The bilinearity of $\langle \cdot, \cdot \rangle_{T_\rho^* M}$ is immediate from the definition. For $\phi \in C_c^\infty(\mathbb{R}^d)$, Assumption 4 implies that $\langle \phi, \phi \rangle_{T_\rho^* M} = 0$ if and only if $\phi = 0$. \square

⁷A pre-Hilbert (or inner-product) space satisfies the usual axioms of a Hilbert space, except for completeness. That is, it does not necessarily contain the limit points of all Cauchy sequences, see [40, Section 3.1].

The cotangent spaces can now be defined as follows:

Definition 3.6 (Cotangent spaces). For $\rho \in M$, we define the cotangent spaces T_ρ^*M to be the completions⁸ of $(C_c^\infty(\mathbb{R}^d), \langle \cdot, \cdot \rangle_{T_\rho^*M})$.

Remark 3.7. On a practical level, the completion construction extends the definition (28) to functions ϕ such that $\langle \phi, \phi \rangle_{T_\rho^*M}$ can be defined⁹ and $\langle \phi, \phi \rangle_{T_\rho^*M} < \infty$. In particular, if $\rho \in M$ and $I_{\text{Stein}}^k(\rho) < \infty$ with I_{Stein}^k defined as in (10), then $\frac{\rho}{\pi} \in T_\rho^*M$, and

$$I_{\text{Stein}}^k(\rho) = \left\| \frac{\rho}{\pi} \right\|_{T_\rho^*M}^2 = \left\| \frac{\delta \text{KL}}{\delta \rho} \right\|_{T_\rho^*M}^2. \quad (30)$$

We will revisit this identity in Section 6.

The next result shows that the Onsager operators \mathbb{K}_ρ naturally translate between the tangent and cotangent spaces, substantiating Remark 3.4:

Proposition 3.8 (Duality). *For any $\rho \in M$, the Onsager operator \mathbb{K}_ρ extends to an isometric isomorphism between the Hilbert spaces T_ρ^*M and $T_\rho M$. That is, the extension (denoted by the same symbol) satisfies*

$$\langle \mathbb{K}_\rho \phi, \mathbb{K}_\rho \psi \rangle_{T_\rho M} = \langle \phi, \psi \rangle_{T_\rho^*M}, \quad (31)$$

for all $\phi, \psi \in T_\rho^*M$.

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

$$\langle \mathbb{K}_\rho \phi, \mathbb{K}_\rho \psi \rangle_{T_\rho M} = \langle \mathcal{T}_{k,\rho} \nabla \phi, \mathcal{T}_{k,\rho} \nabla \psi \rangle_{\mathcal{H}_k^d} = \langle \mathcal{T}_{k,\rho} \nabla \phi, \nabla \psi \rangle_{(L^2(\rho))^d} = \langle \phi, \psi \rangle_{T_\rho^*M}. \quad (32)$$

Here, the first identity follows from the definition (24) and Proposition 2.1(a), while the second identity is implied by the adjoint relation (17). The third identity is a direct consequence of the definition (28). From (32), we see that \mathbb{K}_ρ is a linear isometry from $C_c^\infty(\mathbb{R}^d)$ to $T_\rho M$, and hence can be uniquely extended to an isometry $\widehat{\mathbb{K}}_\rho$ on the completion T_ρ^*M (see [65, Theorem I.7]). Being an isometry, it is clear that $\widehat{\mathbb{K}}_\rho$ is injective. It remains to show that $\widehat{\mathbb{K}}_\rho$ is surjective. To this end, it is sufficient to prove that $\mathbb{K}_\rho(C_c^\infty(\mathbb{R}^d))$ is dense in $T_\rho M$. For this, let us assume to the contrary that $\mathbb{K}_\rho(C_c^\infty(\mathbb{R}^d))$ is not dense. Then there exists $\chi \in T_\rho M$ with $\chi \neq 0$ such that $\langle \chi, \mathbb{K}_\rho \phi \rangle_{T_\rho M} = 0$, for all $\phi \in C_c^\infty(\mathbb{R}^d)$. By Definition 3.1, there exists $v \in \mathcal{H}_k^d$ such that $\chi + \nabla \cdot (\rho v) = 0$ in the sense of distributions, as well as a sequence $(\psi_n) \subset C_c^\infty(\mathbb{R}^d)$ such that $\mathcal{T}_{k,\rho} \nabla \psi_n \rightarrow v$ in \mathcal{H}_k^d . We then see that

$$0 = \langle \chi, \mathbb{K}_\rho \psi_n \rangle_{T_\rho M} = -\langle v, \mathcal{T}_{k,\rho} \nabla \psi_n \rangle_{\mathcal{H}_k^d} \rightarrow -\|v\|_{\mathcal{H}_k^d}^2, \quad (33)$$

implying that $v = 0$. From the second statement in [23, Lemma 7], implied by Proposition 2.3, it then follows that $\chi = 0$, contradicting the assumption $\chi \neq 0$ from before, and hence concluding the proof. \square

We can leverage the correspondence between $T_\rho M$ and T_ρ^*M through \mathbb{K}_ρ provided by Proposition 3.8 to set up an associated duality relation. This duality is natural in that it coincides with the duality between $\mathcal{D}'(\mathbb{R}^d)$ and $C_c^\infty(\mathbb{R}^d)$ whenever both are defined:

Corollary 3.9. *For any $\rho \in M$, we can define the duality relation*

$$T_\rho^*M \langle \phi, \xi \rangle_{T_\rho M} := \langle \mathbb{K}_\rho \phi, \xi \rangle_{T_\rho M}, \quad \phi \in T_\rho^*M, \xi \in T_\rho M. \quad (34)$$

*In particular, T_ρ^*M is a representation of the dual of $T_\rho M$. If $\phi \in C_c^\infty(\mathbb{R}^d) \subset T_\rho^*M$, then we have*

$$T_\rho^*M \langle \phi, \xi \rangle_{T_\rho M} = \mathcal{D}'(\mathbb{R}^d) \langle \xi, \phi \rangle_{C_c^\infty(\mathbb{R}^d)} = \int_{\mathbb{R}^d} v \cdot \nabla \phi \, d\rho, \quad (35)$$

where $\xi + \nabla \cdot (\rho v) = 0$ and $v \in \overline{\mathcal{T}_{k,\rho} \nabla C_c^\infty(\mathbb{R}^d)}^{\mathcal{H}_k^d}$.

⁸Any pre-Hilbert space can be upgraded to a Hilbert space, intuitively by considering all limit points. For a rigorous survey of the completion construction see [40, Section 1.6, Theorem 3.2-3].

⁹Note that this does not require ϕ to be differentiable; indeed (28) can be extended to nondifferentiable ϕ and ψ using integration by parts.

Proof. By Proposition 3.8 and the Riesz representation theorem, (34) establishes a one-to-one correspondence between the topological dual of $T_\rho M$ and $T_\rho^* M$. The second identity in (35) is satisfied by definition, see Remark 3.4. To obtain the first identity, consider a sequence $(\xi_n) \subset T_\rho^* M$ with $\xi_n \rightarrow \xi$, such that there exists a sequence $(\psi_n) \subset C_c^\infty(\mathbb{R}^d)$ satisfying $\xi_n + \nabla \cdot (\rho \mathcal{T}_{k,\rho} \nabla \psi_n) = 0$ in the sense of distributions. Then we have

$$\mathcal{D}'(\mathbb{R}^d) \langle \xi_n, \phi \rangle_{C_c^\infty(\mathbb{R}^d)} = \int_{\mathbb{R}^d} (\mathcal{T}_{k,\rho} \nabla \psi_n) \cdot \nabla \phi \, d\rho = \langle \psi_n, \phi \rangle_{T_\rho^* M} = \langle \mathbb{K}_\rho \psi_n, \mathbb{K}_\rho \phi \rangle_{T_\rho M} \quad (36a)$$

$$= T_\rho^* M \langle \phi, \mathbb{K}_\rho \psi_n \rangle_{T_\rho M} = T_\rho^* M \langle \phi, \xi_n \rangle_{T_\rho M}, \quad (36b)$$

where the first inequality is a consequence of the definition of $\mathcal{D}'(\mathbb{R}^d) \langle \cdot, \cdot \rangle_{C_c^\infty(\mathbb{R}^d)}$, the second equality follows from (28), the third equality follows from (31), and the fourth equality follows from (34). Finally, we obtain (35) by passing to the limit, noting that all operations are continuous. \square

As an consequence of this duality and the Banach-Alaoglu Theorem, we obtain compactness of the (sub-)level sets of the Stein-Fisher information, relevant later in the proof of Theorem 6.1.

Corollary 3.10. *For any $C > 0$, the sets $\{\rho \in M : I_{\text{Stein}}^k(\rho) \leq C\}$ are pre-compact in the topology characterised by the convergence:*

$$\rho^\epsilon \underset{\epsilon \rightarrow 0}{\overset{*}{\rightrightarrows}} \rho \quad \Leftrightarrow \quad T_\pi^* M \left\langle \frac{\rho^\epsilon}{\pi}, \xi \right\rangle_{T_\pi M} \rightarrow T_\pi^* M \left\langle \frac{\rho}{\pi}, \xi \right\rangle_{T_\pi M} \quad \text{for all } \xi \in T_\pi M. \quad (37)$$

We next provide a reformulation of the Stein PDE (8) in terms of an energy-dissipation (in-)equality (see [2, Chapter 11]), using the framework developed in this section:

Proposition 3.11 (Energy-dissipation equality). *For $T > 0$, let $\rho : [0, T] \rightarrow M$ be a curve such that $t \mapsto \text{KL}(\rho_t)$ is differentiable and for all $t \in [0, T]$,*

$$\partial_t \rho_t \in T_{\rho_t} M, \quad \frac{\delta \text{KL}}{\delta \rho}(\rho_t) \in T_{\rho_t}^* M, \quad \text{and} \quad \frac{d}{dt} \text{KL}(\rho_t) = \left\langle \frac{\delta \text{KL}}{\delta \rho}(\rho_t), \partial_t \rho_t \right\rangle_{T_{\rho_t} M}. \quad (38)$$

Then the following statements are equivalent:

the Stein PDE (8) holds for all $t \in [0, T]$,

\Leftrightarrow

$\partial_t \rho_t = -\mathbb{K}_{\rho_t} \frac{\delta \text{KL}}{\delta \rho}(\rho_t)$ for all $t \in [0, T]$,

\Leftrightarrow

$$\text{KL}(\rho_T) - \text{KL}(\rho_0) + \int_0^T \left(\frac{1}{2} \|\partial_t \rho_t\|_{T_{\rho_t} M}^2 + \frac{1}{2} \left\| \frac{\delta \text{KL}}{\delta \rho}(\rho_t) \right\|_{T_{\rho_t}^* M}^2 \right) dt = 0. \quad (39)$$

Moreover, for any curve satisfying (38), the left-hand side of (39) is nonnegative.

Remark 3.12. The assumptions (38) are made for convenience, and we refer to [2] for generalisations. Note that the chain rule, i.e. the last condition of (38), is expected to hold at a formal level, combining (25) and (35). Since (39) is always non-negative, the proposition continues to hold if ‘=’ is replaced by ‘ \leq ’; analogues of (39) are therefore often called *energy-dissipation inequalities* in the literature.

Proof. Take any curve satisfying (38). The statement follows immediately by applying (38) and completing the square:

$$\begin{aligned} & \text{KL}(\rho_T) - \text{KL}(\rho_0) + \frac{1}{2} \int_0^T \left(\|\partial_t \rho_t\|_{T_{\rho_t} M}^2 + \left\| \frac{\delta \text{KL}}{\delta \rho}(\rho_t) \right\|_{T_{\rho_t}^* M}^2 \right) dt \\ &= \int_0^T \left(\left\langle \frac{\delta \text{KL}}{\delta \rho}(\rho_t), \partial_t \rho_t \right\rangle_{T_{\rho_t} M} + \frac{1}{2} \|\partial_t \rho_t\|_{T_{\rho_t} M}^2 + \frac{1}{2} \left\| \frac{\delta \text{KL}}{\delta \rho}(\rho_t) \right\|_{T_{\rho_t}^* M}^2 \right) dt \\ & \stackrel{(31),(34)}{=} \frac{1}{2} \int_0^T \left\| \partial_t \rho_t + \mathbb{K}_{\rho_t} \frac{\delta \text{KL}}{\delta \rho}(\rho_t) \right\|_{T_{\rho_t} M}^2 dt. \end{aligned}$$

\square

We conclude this section with a remark on the relationship between the functional-analytic frameworks associated to the Stein and Wasserstein geometries. For this, we recall that in the Wasserstein geometry, the tangent and cotangent spaces are given by the Sobolev spaces $H^{-1}(\rho)$ and $H^1(\rho)$, respectively, see [54].

Lemma 3.13 (Comparison with the Wasserstein setting). *We have*

$$H^1(\rho) \hookrightarrow T_\rho^*M \quad \text{and} \quad T_\rho M \hookrightarrow H^{-1}(\rho), \quad (40)$$

where \hookrightarrow denotes containment with continuous inclusion.

Remark 3.14. The fact that the tangent spaces $T_\rho M$ in the Stein geometry are contained in the tangent spaces for the Wasserstein geometry is ultimately due to the fact that the movement of the particles is restricted to vector fields belonging to reproducing kernel Hilbert spaces in SVGD.

Proof. For the first statement, it is sufficient to show that there exists a constant $C > 0$ such that $\|\phi\|_{T_\rho^*M} \leq C\|\phi\|_{H^1(\rho)}$, for all $\phi \in C_c^\infty(\mathbb{R}^d)$. This follows immediately from

$$\langle \phi, \phi \rangle_{T_\rho^*M} = \langle \mathcal{T}_{k,\rho} \nabla \phi, \nabla \phi \rangle_{(L^2(\rho))^d} \leq \|\mathcal{T}_{k,\rho}\|_{(L^2(\rho))^d \rightarrow (L^2(\rho))^d} \|\nabla \phi\|_{(L^2(\rho))^d}, \quad \phi \in C_c^\infty(\mathbb{R}^d), \quad (41)$$

noting that $\mathcal{T}_{k,\rho}$ is bounded on $(L^2(\rho))^d$ by Proposition 2.1 and therefore $\|\mathcal{T}_{k,\rho}\|_{(L^2(\rho))^d} < \infty$. The second statement follows immediately by the duality established in Corollary 3.9, see [69, Theorem 4.10]. \square

4 Large deviations corresponding to the mean field limit

In this section we introduce and derive the large-deviation principle for the empirical measure $\rho_t^{(N)}$ associated to the SDE (6) as $N \rightarrow \infty$. The derivation will partly be formal via a standard tilting technique; rigorous results for similar stochastic systems can be found in the classic works [17] and [24, Ch. 13.3].

As mentioned in Subsection 1.1, the (random) path $\rho^{(N)} := (\rho_t^{(N)})_{t \in [0, T]}$ converges weakly as $N \rightarrow \infty$ to the solution $\rho^{(\infty)}$ of the Stein PDE (8). This means that for any continuity set \mathcal{A} of paths [6, Th. 2.1],

$$\mathbb{P}(\rho^{(N)} \in \mathcal{A}) \xrightarrow{N \rightarrow \infty} \delta_{\rho^{(\infty)}}(\mathcal{A}),$$

that is, the probability vanishes for any atypical path $\rho \neq \rho^{(\infty)}$. The large-deviation principle quantifies the exponential rate of this convergence:

$$\mathbb{P}(\rho^{(N)} \in \mathcal{A}) \stackrel{N \rightarrow \infty}{\sim} \exp(-N \inf_{\rho \in \mathcal{A}} \mathcal{I}_{[0, T]}(\rho)), \quad (42)$$

where the *rate functional* $\mathcal{I}_{[0, T]}$ satisfies $\mathcal{I}_{[0, T]}(\rho^{(\infty)}) = 0$ and $\mathcal{I}_{[0, T]}(\rho) > 0$ for any path $\rho \neq \rho^{(\infty)}$. In other words, the magnitude of $\mathcal{I}_{[0, T]}(\rho)$ quantifies the ‘unlikeliness’ of the particular path ρ as a deviation from $\rho^{(\infty)}$, in the exponential scaling indicated above. The infimum on the right-hand side appears because the process will follow the least unlikely path with overwhelming probability; for the precise definition of the large-deviation principle we refer to [19].

Remark 4.1. Throughout this paper we shall formally derive large-deviation principles (42) for small balls $\mathcal{B}_\epsilon(\rho)$ around an unlikely event ρ , i.e.

$$\mathbb{P}(\rho^{(N)} \in \mathcal{B}_\epsilon(\rho)) \stackrel{N \rightarrow \infty}{\sim} \exp(-N \mathcal{I}_{[0, T]}(\rho)).$$

This is a common proof technique for both the large-deviation lower and upper bounds in the rigorous definition, see [19, Sec. 1.2].

Remark 4.2. For brevity we shall largely ignore the role of the initial condition ρ_0 . Implicitly we will always assume that the initial positions X_0^i of all particles are chosen deterministically, in such a way that $\rho_0^{(N)}$ converges weakly to some given ρ_0 . Theorem 6.1 below shows that the leading order contribution as $T \rightarrow \infty$ is independent of ρ_0 .

Our main result provides an expression for the rate functional $\mathcal{I}_{[0, T]}$ in terms of the $T_\rho M$ -norm introduced in Definition 3.1. We postpone a discussion of its interpretation until Sections 6 and 7.

Theorem 4.3 (Large-deviation principle, formal). *The path $(\rho_t^{(N)})_{t \in [0, T]}$ of the empirical measure (5) associated to the SDE (6) satisfies a large-deviation principle (42) with rate functional $\mathcal{I}_{[0, T]}$, given by*

$$\mathcal{I}_{[0, T]}(\rho) = \frac{1}{4} \int_0^T \left\| \partial_t \rho_t - \nabla_x \cdot \left(\rho_t \int_{\mathbb{R}^d} [k(\cdot, y) \nabla V(y) - \nabla_y k(\cdot, y)] \rho_t(dy) \right) \right\|_{T_\rho M}^2 dt, \quad (43)$$

for paths ρ satisfying (38).

We note that the expression (43) can be extended to arbitrary paths $\rho : [0, T] \rightarrow \mathcal{P}(\mathbb{R}^d)$, possibly taking the value infinity; for brevity we will focus on sufficiently regular paths in the sense of (38).

In the remainder of this section, we outline the proof of Theorem 4.3. The key idea is to tilt the underlying probability measure using Girsanov transformations so that the atypical path ρ becomes the typical one for the new, tilted measure. The very same technique is common in importance sampling for diffusions, see [35, 78] and [56, Section 2.2], and sequential Monte Carlo methods [18, 21, 66], used to simulate the occurrence of rare (=atypical) events.

The calculation of the large-deviation rate functional requires the construction of the ‘exponential martingale’¹⁰, for which it will be helpful to know the generator¹¹ of the process $\rho_t^{(N)}$ explicitly.

Lemma 4.4. *For each $N \in \mathbb{N}$, the generator of the Markov process $\rho_t^{(N)}$ defined by (6) and (5) is:*

$$\begin{aligned} (\mathcal{Q}^{(N)}F)(\rho) &= \iint_{\mathbb{R}^d \times \mathbb{R}^d} \left[-k(x, y) \nabla V(y) + \nabla_y k(x, y) \right] \cdot \nabla_x \left(\frac{\delta F}{\delta \rho}(\rho) \right)(x) \rho(dx) \rho(dy) \\ &\quad + \frac{1}{N} \iint_{\mathbb{R}^d \times \mathbb{R}^d} k(x, y) \nabla_x \cdot \nabla_y \left(\frac{\delta^2 F}{\delta \rho^2}(\rho) \right)(x, y) \rho(dx) \rho(dy) \\ &\quad + \frac{1}{N} \int_{\mathbb{R}^d} k(x, x) \Delta \left(\frac{\delta F}{\delta \rho}(\rho) \right)(x) \rho(dx), \end{aligned} \quad (44)$$

where $F : \mathcal{P}(\mathbb{R}^d) \rightarrow \mathbb{R}$ is a test function of sufficient regularity.

The proof follows a standard calculation involving Itô’s formula that we postpone to the appendix. The following result shows that the process can be perturbed or *tilted* by adding an additional, time-dependent drift such that the Radon-Nikodym derivative is explicit.

Lemma 4.5 (Girsanov transformation). *Let $\mathbb{P}_{[0, T]}^{(N)}$ be the law of the empirical measure process $(\rho_t^{(N)})_{t \in [0, T]}$ associated to the SDE (6), fix a test function $G : [0, T] \times \mathcal{P}(\mathbb{R}^d)$ of sufficient regularity¹², and define the tilted measure $\mathbb{P}_{[0, T]}^{(N, G)}$ through the Radon-Nikodym derivative,*

$$\frac{d\mathbb{P}_{[0, T]}^{(N, G)}}{d\mathbb{P}_{[0, T]}^{(N)}}(\rho) = \exp \left(NG_T(\rho_T) - NG_0(\rho_0) - N \int_0^T (\partial_t G_t)(\rho_t) dt - N \int_0^T (\mathcal{H}^{(N)} G_t)(\rho_t) dt \right), \quad (45)$$

where the operator $\mathcal{H}^{(N)}$ is defined as

$$(\mathcal{H}^{(N)} G)(\rho) := \frac{1}{N} e^{-NG(\rho)} (\mathcal{Q}^{(N)} e^{NG(\rho)})(\rho). \quad (46)$$

Then $\mathbb{P}_{[0, T]}^{(N, G)}$ is the law of the (time-inhomogeneous) Markov process with generator

$$(\mathcal{Q}_t^{(N, G)} F)(\rho) := (\mathcal{Q}_t^{(N)} F)(\rho) + \iint_{\mathbb{R}^d \times \mathbb{R}^d} k(x, y) \nabla_x \cdot \nabla_y \left(\frac{\delta F}{\delta \rho}(\rho)(x) \frac{\delta G_t}{\delta \rho}(\rho)(y) + \frac{\delta G_t}{\delta \rho}(\rho)(x) \frac{\delta F}{\delta \rho}(\rho)(y) \right) \rho(dx) \rho(dy). \quad (47)$$

Remark 4.6. The limit of the operator $\mathcal{H}^{(N)}$ coincides with the generator of the nonlinear Nisio semigroup in the framework of [24].

Proof. The statement follows from the general Girsanov transformation formula [63, Th. 4.2], adapted to allow for time-dependent test functions G , see also [38, Th. A1.7.3] and [24, Sec. 8.6.1.1]:

$$(\mathcal{Q}_t^{(N, G)} F)(\rho) = e^{-NG_t(\rho)} (\mathcal{Q}_t^{(N)} F e^{NG_t(\rho)})(\rho) - e^{-NG_t(\rho)} F(\rho) (\mathcal{Q}_t^{(N)} e^{NG_t(\rho)})(\rho).$$

□

In the following we explicitly calculate (46) and pass to the limit as $N \rightarrow \infty$.

¹⁰The exponential martingale is the right-hand side of (45) applied to the random process $\rho_t^{(N)}$.

¹¹Since the process $\rho_t^{(N)}$ takes (random) values in $\mathcal{P}(\mathbb{R}^d)$, its generator acts on functionals $F : \mathcal{P}(\mathbb{R}^d) \rightarrow \mathbb{R}$ of sufficient regularity. For background on measure-valued stochastic processes we refer the reader to [16].

¹²A convenient class of test functions is given by $G \in C_b^2(0, T; C_b^2(\mathcal{P}(\mathbb{R}^d)))$, where the derivative in $\mathcal{P}(\mathbb{R}^d)$ is understood in the sense of (25). See [79, Theorem 2.2.1] for the general Novikov condition (in finite dimensions) and [63] for an even more general condition.

Lemma 4.7. For any test function $G : [0, T] \times \mathcal{P}(\mathbb{R}^d)$ of sufficient regularity (as above), we have

$$(\mathcal{H}^{(N)} G)(\rho) \xrightarrow{N \rightarrow \infty} \mathcal{H}\left(\rho, \frac{\delta G}{\delta \rho}(\rho)\right) := -\left\langle \frac{\delta G}{\delta \rho}(\rho), \frac{\delta \text{KL}}{\delta \rho}(\rho) \right\rangle_{T_\rho^* M} + \left\| \frac{\delta G}{\delta \rho}(\rho) \right\|_{T_\rho^* M}^2,$$

where $\langle \cdot, \cdot \rangle_{T_\rho^* M}$ is defined in (28).

We postpone this calculation to the appendix.

Lemma 4.8 (Mean-field limit, formal). Fix a test function $G : [0, T] \times \mathcal{P}(\mathbb{R}^d)$ of sufficient regularity (as above) and let $\rho_t^{(N, G)}$ be the process with generator (47). Then $\rho^{(N, G)} \rightarrow \rho^{(\infty, G)}$, which weakly solves the “tilted Stein PDE”:

$$\partial_t \rho_t(x) = \nabla_x \cdot \left(\rho_t(x) \int_{\mathbb{R}^d} \left[k(x, y) \nabla V(y) - \nabla_y k(x, y) - 2k(x, y) \nabla_y \frac{\delta G_t}{\delta \rho}(\rho_t)(y) \right] \rho_t(dy) \right). \quad (48)$$

Proof sketch. Clearly the generator (47) converges pointwise in ρ to

$$\begin{aligned} (\mathcal{Q}_t^{(\infty, G)} F)(\rho) &:= \iint_{\mathbb{R}^d \times \mathbb{R}^d} \left[-k(x, y) \nabla V(y) + \nabla_y k(x, y) \right] \cdot \nabla_x \left(\frac{\delta F}{\delta \rho}(\rho) \right)(x) \rho(dx) \rho(dy) \\ &\quad + \iint_{\mathbb{R}^d \times \mathbb{R}^d} k(x, y) \nabla_x \cdot \nabla_y \left(\frac{\delta F}{\delta \rho}(\rho) \right)(x) \frac{\delta G_t}{\delta \rho}(\rho)(y) + \frac{\delta G_t}{\delta \rho}(\rho)(x) \frac{\delta F}{\delta \rho}(\rho)(y) \rho(dx) \rho(dy). \end{aligned}$$

Hence formally by [42, Th. 2.12], the process $\rho_t^{(N, G)}$ converges to some (a priori stochastic) process $\rho_t^{(\infty, G)}$. It remains to show that this process satisfies (48) and is thus deterministic.

Let $P_t^{(\infty, G)}$ be the time marginal of the path measure $\mathbb{P}_{[0, T]}^{(\infty, G)}$ and let us make the ansatz that it is indeed deterministic: $P_t^{(\infty, G)} = \delta_{\rho_t}$ for some (by assumption) sufficiently regular path $\rho_t(dx) = \rho_t(x) dx$. Then using the Chapman-Kolmogorov forward equation,

$$\begin{aligned} \int_{\mathbb{R}^d} \frac{\delta F}{\delta \rho_t}(\rho_t)(x) \partial_t \rho_t(dx) &= \frac{d}{dt} F(\rho_t) = \frac{d}{dt} \int F(\rho) P_t(d\rho) = \int (\mathcal{Q}_t^{(\infty, G)} F)(\rho) P_t(d\rho) \\ &= (\mathcal{Q}_t^{(\infty, G)} F)(\rho_t) = \int_{\mathbb{R}^d} \frac{\delta F}{\delta \rho}(\rho_t)(x) B^{(G_t)}(\rho_t(x)) dx, \end{aligned}$$

where $B^{(G_t)}(\rho_t(x))$ is the right-hand side of (47). Since this equation holds for arbitrary (sufficiently regular) test functions F , it follows that $\partial_t \rho_t = B^{(G_t)}(\rho_t)$ weakly, and so the ansatz is justified for $\rho_t = \rho_t^{(\infty, G)}$. \square

Remark 4.9. In the notation introduced in Section 3, the tilted Stein PDE (48) becomes

$$\partial_t \rho_t = -\mathbb{K}_{\rho_t} \frac{\delta \text{KL}}{\delta \rho}(\rho_t) + 2\mathbb{K}_{\rho_t} \frac{\delta G_t}{\delta \rho}(\rho_t).$$

We finally have all the ingredients to prove the main result of this section.

Proof sketch of Theorem 4.3. To simplify, we only derive the rate functional for an arbitrary path $\rho = (\rho_t)_{t \in [0, T]}$ satisfying (38). Corresponding to this path, let $\xi_t \in T_{\rho_t}^* M$ be maximal in

$$T_{\rho_t}^* M \langle \xi_t, \partial_t \rho_t \rangle_{T_{\rho_t}^* M} - \mathcal{H}(\rho_t, \xi_t) = \left\langle \xi_t, \partial_t \rho_t + \mathbb{K}_{\rho_t} \frac{\delta \text{KL}}{\delta \rho}(\rho_t) \right\rangle_{T_{\rho_t}^* M} + T_{\rho_t}^* M \langle \xi_t, \mathbb{K}_{\rho_t} \xi_t \rangle_{T_{\rho_t}^* M},$$

pointwise in $t \in [0, T]$, where the brackets are defined in Corollary 3.9, and \mathcal{H} is the limit obtained in Lemma 4.7. Again for simplicity we shall assume that this maximiser exists, and in fact $\xi_t \in C_c^\infty(\mathbb{R}^d) \subset T_{\rho_t}^* M$.

Upon differentiation with respect to ξ_t we recover the tilted Stein PDE (48) with $G_t(\rho) := \mathcal{D}'(\mathbb{R}^d) \langle \rho_t, \xi_t \rangle_{C_c^\infty(\mathbb{R}^d)}$ so that $(\delta G_t / \delta \rho)(\rho_t) = \xi_t$.

Thus for this particular choice, Lemma 4.8 shows that the tilted process converges to the path we picked in the beginning of the proof, i.e.

$$\mathbb{P}_{[0, T]}^{(N, G)} \rightarrow \delta_\rho. \quad (49)$$

Now pick an arbitrary small ball $\mathcal{B}_\epsilon(\rho)$ (in Skorokhod space) around the path ρ . By Lemma 4.5 we may change the measure and write, for small $\epsilon > 0$:

$$\frac{1}{N} \log \mathbb{P}_{[0, T]}^{(N)}(\mathcal{B}_\epsilon(\rho)) = \frac{1}{N} \log \int_{\mathcal{B}_\epsilon(\rho)} \frac{d\mathbb{P}_{[0, T]}^{(N)}}{d\mathbb{P}_{[0, T]}^{(N, G)}}(\hat{\rho}) d\mathbb{P}_{[0, T]}^{(N, G)}(d\hat{\rho}) \approx \frac{1}{N} \log \frac{d\mathbb{P}_{[0, T]}^{(N)}}{d\mathbb{P}_{[0, T]}^{(N, G)}}(\rho) + \frac{1}{N} \log \mathbb{P}_{[0, T]}^{(N, G)}(\mathcal{B}_\epsilon(\rho)). \quad (50)$$

By (49) the last term vanishes, and so by Lemmas 4.5 and 4.7 we find as $N \rightarrow \infty$ and small $\epsilon > 0$,

$$\begin{aligned}
\frac{1}{N} \log \mathbb{P}_{[0,T]}^{(N)}(\mathcal{B}_\epsilon(\rho)) &\rightarrow -\mathcal{D}'(\mathbb{R}^d) \langle \rho_T, \xi_T \rangle_{C_c^\infty(\mathbb{R}^d)} + \mathcal{D}'(\mathbb{R}^d) \langle \rho_0, \xi_0 \rangle_{C_c^\infty(\mathbb{R}^d)} - \int_0^T \mathcal{D}'(\mathbb{R}^d) \langle \rho_t, \partial_t \xi_t \rangle_{C_c^\infty(\mathbb{R}^d)} dt - \int_0^T \mathcal{H}(\rho_t, \xi_t) dt \\
&= - \sup_{\hat{\xi}: (0,T) \rightarrow T_\rho^* M} \int_0^T [T_{\rho_t}^* M \langle \hat{\xi}_t, \partial_t \rho_t \rangle_{T_{\rho_t} M} - \mathcal{H}(\rho_t, \hat{\xi}_t)] dt \\
&= - \sup_{\hat{\xi}: (0,T) \rightarrow T_\rho^* M} \int_0^T \left[T_{\rho_t}^* M \left\langle \hat{\xi}_t, \partial_t \rho_t + \mathbb{K}_{\rho_t} \frac{\delta \text{KL}}{\delta \rho}(\rho_t) \right\rangle_{T_{\rho_t} M} - \|\hat{\xi}_t\|_{T_\rho^* M}^2 \right] dt \\
&= -\mathcal{I}_{[0,T]}(\rho),
\end{aligned}$$

where the supremum over tiltings $\hat{\xi}$ appears due to the definition of ξ .

In general, the perturbation functions ξ and G do not have sufficient regularity to apply Lemma 4.5, if they exist at all, so one typically needs technically demanding approximation arguments to make this into a rigorous argument, see for example [24, Ch. 13] and [17]. \square

From the large-deviation result in Theorem 4.3 and the contraction principle [19, Th. 4.2.1] we immediately obtain the large-deviation principle for the ergodic limit (3). The ensuing rate functional will be further analysed in Section 6.

Corollary 4.10. *Fix $T > 0$ and $\rho_0 \in M$. Let $(\rho_t^{(N)})_{t \in [0,T]}$ be the path of the empirical measure associated to the SDE (6), and let*

$$\bar{\rho}_T^{(N)} := \frac{1}{T} \int_0^T \rho_t^{(N)} dt$$

be the ergodic average thereof. Then $\bar{\rho}_T^{(N)}$ satisfies a large-deviation principle as $N \rightarrow \infty$, i.e.

$$\mathbb{P}(\bar{\rho}_T^{(N)} \in \mathcal{A}) \stackrel{N \rightarrow \infty}{\sim} \exp(-N \inf_{\bar{\rho} \in \mathcal{A}} \bar{\mathcal{I}}_T(\bar{\rho})),$$

with rate functional:

$$\bar{\mathcal{I}}_T(\bar{\rho}) := \inf_{\substack{\hat{\rho}: [0,T] \rightarrow M: \\ \hat{\rho}_0 = \rho_0, \\ T^{-1} \int_0^T \hat{\rho}_t dt = \bar{\rho}}} \mathcal{I}_{[0,T]}(\hat{\rho}). \quad (51)$$

5 Connecting gradient flows to large deviations

As stated in the Introduction, any evolution equation of gradient flow type in fact admits many other non-equivalent gradient flow structures [20]. In the case of the Stein PDE (8) this phenomenon is exemplified by the structures proposed in [44, 23] and [13]. However, each gradient flow structure is related to a particular form of the noise in the corresponding interacting particle system. In this section we leverage our results from Sections 3 and 4 to make our Informal Result 1.1 precise: the gradient flow structure from Section 3 corresponds to the noise described by the SDE (6). Our rigorous statement draws a connection between the reformulation of the gradient flow dynamics in terms of the energy-dissipation (in-)equality (39) and the large-deviation functional (43):

Theorem 5.1 (Connection between energy-dissipation and large deviations). *For any curve $\rho : [0, T] \rightarrow M$ such that $t \mapsto \text{KL}(\rho_t)$ is differentiable and (38) holds for all $t \in [0, T]$, the left-hand side of (39) coincides with $\mathcal{I}_{[0,T]}(\rho)$ up to a factor of $\frac{1}{2}$, that is*

$$\mathcal{I}_{[0,T]}(\rho) = \frac{1}{2} \text{KL}(\rho_T) - \frac{1}{2} \text{KL}(\rho_0) + \frac{1}{4} \int_0^T \|\partial_t \rho_t\|_{T_{\rho_t} M}^2 dt + \int_0^T \left\| \frac{\delta \frac{1}{2} \text{KL}}{\delta \rho}(\rho_t) \right\|_{T_{\rho_t} M}^2 dt. \quad (52)$$

This implies that the large deviations from Theorem 4.3 uniquely induce the gradient flow system with driving energy $\frac{1}{2} \text{KL}(\rho)$ and cotangent norm $\|\cdot\|_{T_\rho^* M}$ in the sense of [52].

Connections between energy-dissipation and large deviations have a long history in physics, starting from the idea that for non-evolving random systems, the Boltzmann-Gibbs-Helmholtz free energy $\mathcal{F}(\rho)$ of a macroscopic state ρ is related to the probability of corresponding microstates through $\mathbb{P}^{(N)}(B_\epsilon(\rho)) \sim \exp(-N\mathcal{F}(\rho)/(\kappa_B T))$. For the sake of brevity we ignore the Boltzmann constant κ_B and the constant temperature T . A dynamical version of this principle was proposed by Onsager and Machlup [57, 58], showing that for a number of physical examples with reversible randomness on the microscopic level, the path measures behave like

$$\mathbb{P}^{(N)}(\mathcal{B}_\epsilon(\rho)) \sim \exp\left(-N\left[\mathcal{F}(\rho_T) - \mathcal{F}(\rho_0) + \frac{1}{2} \int_0^T |\partial_t \rho_t|_{\rho_t}^2 dt + \frac{1}{2} \int_0^T \left|\frac{\delta \mathcal{F}}{\delta \rho}(\rho_t)\right|_{\rho_t^*}^2 dt\right]\right), \quad (53)$$

at least close to equilibrium. In the above display, \mathcal{F} stands for an appropriate free energy functional, $|\cdot|_{\rho_t}$ and $|\cdot|_{\rho_t^*}$ for suitable dual norms, and $\epsilon > 0$ is assumed to be small. Moreover, Onsager and Machlup demonstrated that these constituents define a corresponding gradient flow structure¹³. Note that the exponent in (53) has the dimensions of a free energy (ignoring the Boltzmann constant and the constant temperature), which is consistent with the Boltzmann-Gibbs-Helmholtz free energy as described above.

More recently, this principle was extended to include more general dynamics that are also allowed to evolve far away from their equilibrium state [54]. It turns out that for any microscopic *reversible* Markov process, the corresponding large-deviations rate can be decomposed in such a way that it uniquely defines the free energy functional \mathcal{F} and the dissipation mechanism (in (53) encoded in the two norms $|\cdot|_{\rho_t}$ and $|\cdot|_{\rho_t^*}$) of a gradient flow. For quadratic rate functionals, as in our case (43), this decomposition corresponds to an expansion of squares, which basically amounts to connecting the energy-dissipation (in-)equality (3.11) to the large-deviation functional (43). This connection is the rigorous statement of the Onsager-Machlup principle described above, as well as of our Informal Result 1.1.

Proof of Theorem 5.1. The decomposition follows the same argument as the proof of Proposition 3.11. Note that by (31) and (34), the two squared norms are convex duals to each other, i.e. for all $\rho \in M$, $\xi \in T_\rho M$ and $\phi \in T_\rho^* M$

$$\frac{1}{4} \|\xi\|_{T_\rho M}^2 = \sup_{\phi \in T_\rho^* M} T_\rho^* M \langle \phi, \xi \rangle_{T_\rho M} - \|\phi\|_{T_\rho^* M}^2, \quad \text{and} \quad \|\phi\|_{T_\rho^* M}^2 = \sup_{\xi \in T_\rho M} T_\rho^* M \langle \phi, \xi \rangle_{T_\rho M} - \frac{1}{4} \|\xi\|_{T_\rho M}^2.$$

This indeed implies that (52) is a decomposition in the sense of [52, eq. (1.10)]. The uniqueness of the driving energy $\frac{1}{2}$ KL and cotangent norm $\|\cdot\|_{T_\rho^* M}$ follows from [52, Th. 2.1(ii)]. \square

Remark 5.2. Strictly speaking, this result yields a different gradient flow structure:

$$\partial_t \rho_t = -(2\mathbb{K}_{\rho_t})\left(\frac{1}{2} \text{KL}(\rho_t)\right).$$

The constant 1/2 in front of the Kullback-Leibner divergence is a known issue; it arises because the Kullback-Leibner divergence is related to the difference of large-deviation costs of moving forward and backward in time (note the time-reversal symmetry (57)), hence when only moving forward in time, the constant 1/2 appears. Similarly, the constant 2 in front of the Onsager operator \mathbb{K}_ρ appears as the derivative of the norm $\|\cdot\|_{T_\rho^* M}^2$. We again refer to [52] for the details. Of course, one can also absorb the constant 1/2 in the Onsager operator as we do.

From Theorem 5.1 we immediately obtain the following relation between the Stein-Fisher information and free energy dissipation:

Corollary 5.3 ([23, 49]). *For the solutions ρ of the Stein PDE (8):*

$$\frac{d}{dt} \text{KL}(\rho_t) = -I_{\text{Stein}}^k(\rho_t).$$

Hence the Fisher information controls the convergence for the Stein PDE as $t \rightarrow \infty$ (see also the discussion of the Stein log-Sobolev inequality in [23, Remark 35] and [39]). In the next section we show that the Fisher information also controls the convergence for the stochastic SVGD scheme as both $N \rightarrow \infty$ and $T \rightarrow \infty$.

¹³The reversibility of a Markov process is often called detailed balance in the physics literature to distinguish it from thermodynamical reversibility and was referred to as reciprocity relations by Onsager and Machlup [57]. Moreover, they called the energy-dissipation inequality (39) the principle of least dissipation.

6 Long-time behaviour and the Stein-Fisher information

Generally speaking, large values of rate functionals promise fast convergence, as the corresponding fluctuations are suppressed. To obtain interpretable information from (43), we study the rate functional $\bar{\mathcal{I}}_T$ governing the ergodic average (see Corollary 4.10) in the regime where the final time T is large. As mentioned in the introduction, the leading order term will be given by the Stein-Fisher information (10) (or the kernelised Stein discrepancy). We first show this relation between the Stein-Fisher information and the large-deviation functional, and then investigate the Stein-Fisher information for different kernels.

6.1 From large deviations to the Stein-Fisher information

Recall the large-deviation principle for the ergodic average from Corollary 4.10, for a fixed final time T . By the energy-dissipation decomposition (52) we may write, using a change of variables,

$$\bar{\mathcal{I}}_T(\bar{\rho}) = \inf_{\substack{\hat{\rho}: [0,1] \rightarrow M: \\ \hat{\rho}_0 = \rho_0, \\ \int_0^1 \hat{\rho}_t dt = \bar{\rho}}} \left\{ \frac{1}{2} \text{KL}(\hat{\rho}_1) - \frac{1}{2} \text{KL}(\rho_0) + \frac{1}{4T} \int_0^1 \|\partial_t \hat{\rho}_t\|_{T_{\hat{\rho}_t} M}^2 dt + \frac{T}{4} \int_0^1 \left\| \frac{\delta \text{KL}}{\delta \rho}(\hat{\rho}_t) \right\|_{T_{\hat{\rho}_t}^* M}^2 dt \right\}.$$

Therefore, at least formally, we see that the last term, representing the Stein-Fisher information (see Remark 3.7), becomes dominant and of order $\mathcal{O}(T)$. To make this into a rigorous statement, one might naively take the pointwise limit of $T^{-1} \bar{\mathcal{I}}_T(\bar{\rho})$; however generally this limit does not exist, nor is it the right limit concept to use. To be consistent with the notion of large deviations we will need to use the concept of Γ -convergence [9]. Together, the large-deviation principle and the Γ -convergence will then imply a joint large-deviation principle in N and T , see, for example [5, Sec. 4]. This will be the content of Corollary 6.2.

Let us stress here that the notion of Γ -convergence requires a topology on the underlying space, and that the most natural topology is the one for which the limit $4^{-1} I_k^{\text{Stein}}$ has compact (sub-)level sets, see for example [19, Sec. 1.2] and [9, Lem. 6.2]. In our case, this means that we will choose the topology defined by (37).

Theorem 6.1. *Fix the initial condition ρ_0 such that $\text{KL}(\rho_0) < \infty$. Then in the topology of (37),*

$$\Gamma\text{-}\lim_{T \rightarrow \infty} \frac{1}{T} \bar{\mathcal{I}}_T = \frac{1}{4} I_{\text{Stein}}^k,$$

meaning that

1. for all converging sequences of probability measures $\bar{\rho}_T \xrightarrow[T \rightarrow \infty]{*\pi} \bar{\rho}$,

$$\liminf_{T \rightarrow \infty} \frac{1}{T} \bar{\mathcal{I}}_T(\bar{\rho}_T) \geq \frac{1}{4} I_{\text{Stein}}^k(\bar{\rho}), \quad \text{and} \quad (54)$$

2. for all $\bar{\rho} \in M$, there exists a converging sequence of probability measures $\bar{\rho}_T \xrightarrow[T \rightarrow \infty]{*\pi} \bar{\rho}$ such that

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \bar{\mathcal{I}}_T(\bar{\rho}_T) \leq \frac{1}{4} I_{\text{Stein}}^k(\bar{\rho}). \quad (55)$$

Proof. For the upper bound we take an arbitrary $\bar{\rho}$, for now assuming that $\text{KL}(\bar{\rho}) < \infty$. The statement (55) would be trivial if we could replace the infimum in (51) by the constant path $\bar{\rho}_t \equiv \bar{\rho}$. However this is likely to violate the initial condition, and so we first need to construct a finite-time and finite-cost connecting path between ρ_0 and $\bar{\rho}$. For this construction we shall need two ingredients. The first ingredient is the fact that KL is the ‘quasipotential’, i.e. for all $\hat{\rho} \in M$,

$$\lim_{T \rightarrow \infty} \inf_{\substack{\rho: [0, T] \rightarrow M: \\ \rho_0 = \pi, \rho_T = \hat{\rho}}} \mathcal{I}_{[0, T]}(\rho) = \text{KL}(\hat{\rho}). \quad (56)$$

This statement is standard and can be proven by solving the corresponding Hamilton-Jacobi-Bellman equation, see for example [26]. The second ingredient is the so-called ‘time-reversal symmetry’, meaning that for arbitrary $T > 0$, path $\rho: [0, T] \rightarrow M$ and reversed path $\overleftarrow{\rho}_t := \rho_{T-t}$,

$$\mathcal{I}_{[0, T]}(\rho) - \mathcal{I}_{[0, T]}(\overleftarrow{\rho}) = \text{KL}(\rho_T) - \text{KL}(\rho_0). \quad (57)$$

This symmetry is implied by the reversibility of the process $\rho_t^{(N)}$ [52, Th. 3.3], but it can also be seen directly from the decomposition (52).

We now use these two ingredients to construct a connecting path between ρ_0 and $\bar{\rho}$. By (56), there exists a $T_1 < \infty$ and a path $\rho^1 : [0, T_1] \rightarrow M$ connecting π to ρ_0 so that $\mathcal{I}_{[0, T_1]}(\rho^1) \leq \text{KL}(\rho_0) + 1$. By the time-reversal symmetry (57), the reversal $\overleftarrow{\rho^1}$ of this path connects ρ_0 to π , and satisfies $\mathcal{I}_{[0, T_1]}(\overleftarrow{\rho^1}) \leq 1$. Similarly, there exists a $T_2 < \infty$ and a path $\rho^2 : [0, T_2] \rightarrow M$ connecting π to $\bar{\rho}$ such that $\mathcal{I}_{[0, T_2]}(\rho^2) \leq \text{KL}(\bar{\rho}) + 1$.

From these two paths we construct a new, continuous path $\rho : [0, T] \rightarrow M$ for arbitrary large $T > 0$:

$$\rho_t := \begin{cases} \overleftarrow{\rho_t^1}, & t \in [0, T_1), \\ \rho_{t-T_1}^2, & t \in [T_1, T_1 + T_2), \\ \bar{\rho}, & t \in [T_1 + T_2, T]. \end{cases}$$

This path has the average value

$$\bar{\rho}_T := \frac{1}{T} \int_0^T \rho_t dt = \frac{T_1}{T} \left(\frac{1}{T} \int_0^T \rho_t^1 dt \right) + \frac{T_2}{T} \left(\frac{1}{T} \int_0^T \rho_t^2 dt \right) + \frac{T - T_1 - T_2}{T} \bar{\rho},$$

which clearly converges as claimed, $\bar{\rho}_T \xrightarrow{*} \bar{\rho}$.

Plugging this path and average value into definitions (43), (51) and using (31) yields

$$\begin{aligned} \frac{1}{T} \bar{\mathcal{I}}_T(\bar{\rho}_T) &\leq \frac{1}{T} \mathcal{I}_{[0, T]}(\rho) = \frac{1}{T} \mathcal{I}_{[0, T_1]}(\overleftarrow{\rho^1}) + \frac{1}{T} \mathcal{I}_{[T_1, T_1 + T_2]}(\rho^2) + \frac{1}{T} \mathcal{I}_{[T_1 + T_2, T]}(\bar{\rho}) \\ &\leq \frac{1}{T} + \frac{\text{KL}(\bar{\rho}) + 1}{T} + \frac{T - T_1 - T_2}{T} \left\| \frac{\delta_{\frac{1}{2}} \text{KL}}{\delta \rho}(\bar{\rho}) \right\|_{T \bar{\rho}^* M}^2, \end{aligned}$$

and the upper bound (55) follows by letting $T \rightarrow \infty$ together with the assumption $\text{KL}(\bar{\rho}) < \infty$.

We now handle the case when $\text{KL}(\bar{\rho}) = \infty$ using an additional approximation and a diagonal argument. Without loss of generality we may assume that $I_{\text{Stein}}^k(\bar{\rho}) < \infty$, else the statement (55) would be trivial. It follows from Definition 3.6 and Remark 3.7 that there exists a sequence $\phi^\epsilon \in C_c^\infty(\mathbb{R}^d)$ so that

$$I_{\text{Stein}}^k(\bar{\rho}) = \left\| \frac{\bar{\rho}}{\pi} \right\|_{T \bar{\rho}^* M}^2 \leftarrow \|\phi^\epsilon\|_{T \bar{\rho}^* M}^2 = I_{\text{Stein}}^k(\pi \phi^\epsilon).$$

Without loss of generality we may assume that $\pi \phi^\epsilon$ is a probability measure. Of course this sequence has uniformly bounded Fisher information, so that it has a convergent subsequence by Lemma 3.10. Let us relabel this sequence so that $\pi \phi^\epsilon \xrightarrow{*} \bar{\rho}$. Clearly $\text{KL}(\pi \phi^\epsilon) < \infty$ and so by the construction above there exists an approximating sequence $\bar{\rho}_T^\epsilon \rightarrow \pi \phi^\epsilon$ for which $\limsup_{T \rightarrow \infty} \frac{1}{T} \bar{\mathcal{I}}_T(\bar{\rho}_T^\epsilon) \leq \frac{1}{4} I_{\text{Stein}}^k(\pi \phi^\epsilon)$. We can then define $\bar{\rho}_T := \bar{\rho}_T^{\epsilon_T}$ where we pick $\epsilon_T \rightarrow 0$ sufficiently slowly so that $\bar{\rho}_T \xrightarrow{*} \bar{\rho}$ and

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \bar{\mathcal{I}}_T(\bar{\rho}_T) = \limsup_{\epsilon \rightarrow 0} \limsup_{T \rightarrow \infty} \frac{1}{T} \bar{\mathcal{I}}_T(\bar{\rho}_T^\epsilon) \leq \limsup_{\epsilon \rightarrow 0} \frac{1}{4} I_{\text{Stein}}^k(\pi \phi^\epsilon) = \frac{1}{4} I_{\text{Stein}}^k(\bar{\rho}).$$

For the lower bound (54), pick an arbitrary convergent sequence $\bar{\rho}_T \rightarrow \bar{\rho}$, and for each $T > 0$ an arbitrary path $\rho : [0, T] \rightarrow M$ starting from ρ_0 and with average value $T^{-1} \int_0^T \rho_t dt = \bar{\rho}_T$. We again use the decomposition (52) as well as (30), and neglecting some non-negative terms to derive

$$\frac{1}{T} \mathcal{I}_{[0, T]}(\rho) \geq -\frac{1}{2T} \text{KL}(\rho_0) + \frac{1}{4T} \int_0^T I_{\text{Stein}}^k(\rho_t) dt \geq -\frac{1}{2T} \text{KL}(\rho_0) + \frac{1}{4} I_{\text{Stein}}^k(\bar{\rho}_T),$$

using Jensen's inequality and the fact that the Stein-Fisher information I_{Stein}^k is convex. By taking the infimum over all such paths we find $\frac{1}{T} \bar{\mathcal{I}}_T(\bar{\rho}_T) \geq -\frac{1}{2T} \text{KL}(\rho_0) + \frac{1}{4} I_{\text{Stein}}^k(\bar{\rho}_T)$. Then the lower bound (54) follows from the lower semicontinuity of KL as a consequence of Corollary 3.10. \square

The following result is the mathematically precise statement of our Informal Result 1.2.

Corollary 6.2. *The ergodic average empirical measure $\bar{\rho}_T^{(N)}$ associated to the SDE (6) satisfies the large-deviation principle as first $N \rightarrow \infty$ and then $T \rightarrow \infty$ with rate functional $\frac{1}{4} I_{\text{Stein}}^k$, i.e.*

$$\mathbb{P}^{(N)}(\bar{\rho}_T^{(N)} \in \mathcal{B}_\epsilon(\bar{\rho})) \sim \exp\left(-\frac{1}{4} N T I_{\text{Stein}}^k(\bar{\rho})\right).$$

6.2 Comparing the Stein-Fisher information for different kernels

Corollary 6.2 motivates using the Stein-Fisher information I_{Stein}^k for a principled choice of the kernel k (greater values of I_{Stein}^k promise faster convergence). As stated in Proposition 1.3, the comparison between $I_{\text{Stein}}^{k_1}$ and $I_{\text{Stein}}^{k_2}$ can be made on the basis of the RKHSs \mathcal{H}_{k_1} and \mathcal{H}_{k_2} . Here we provide the proof based on the duality relations established in Section 3.2.

Proof of Proposition 1.3. In this proof, we use the notation $T_\rho M_1$ and $T_\rho M_2$ to distinguish the tangent spaces induced by k_1 and k_2 , respectively, and employ a similar convention for the cotangent spaces. We first show that 2.) implies 1.): By Remark 3.7, it is sufficient to show that $T_\rho^* M_2 \subset T_\rho^* M_1$, with

$$\|\phi\|_{T_\rho^* M_1}^2 \leq \|\phi\|_{T_\rho^* M_2}^2,$$

for all $\phi \in T_\rho^* M_2$. Now, for $\phi \in C_c^\infty(\mathbb{R}^d)$, $\rho \in M$, and $i \in \{1, 2\}$, we see that

$$\|\phi\|_{T_\rho^* M_i} = \sup_{0 \neq \psi \in T_\rho M_i} \frac{T_\rho^* M_i \langle \phi, \psi \rangle_{T_\rho M_i}}{\|\psi\|_{T_\rho M_i}} = \sup_{v \in \mathcal{T}_{k_i, \rho} \nabla C_c^\infty(\mathbb{R}^d)^{\mathcal{H}_{k_i}^d}} \frac{\int_{\mathbb{R}^d} v \cdot \nabla \phi \, d\rho}{\|v\|_{\mathcal{H}_{k_i}^d}} = \sup_{v \in \mathcal{H}_{k_i}^d} \frac{\int_{\mathbb{R}^d} v \cdot \nabla \phi \, d\rho}{\|v\|_{\mathcal{H}_{k_i}^d}}, \quad (58)$$

where the first equality follows from the duality between $T_\rho M_i$ and $T_\rho^* M_i$, the second equality follows directly from Definition 3.1, and the third equality is a consequence of the Helmholtz decomposition in Proposition 2.3. The claim now follows from the fact that by construction, $C_c^\infty(\mathbb{R}^d)$ is dense in $T_\rho^* M_1$ and $T_\rho^* M_2$.

To show that 1.) implies 2.), assume that $v = \mathcal{T}_{k, \rho} \nabla \phi$ and $\xi + \nabla \cdot (\rho v) = 0$. We then have

$$\|v\|_{\mathcal{H}_k^d} = \|\xi\|_{T_\rho M} = \sup_{\psi \in T_\rho^* M} \frac{T_\rho^* M_i \langle \xi, \psi \rangle_{T_\rho M_i}}{\|\psi\|_{T_\rho^* M_i}} = \sup_{\psi \in T_\rho^* M} \frac{\int_{\mathbb{R}^d} v \cdot \nabla \psi \, d\rho}{\|\psi\|_{T_\rho^* M}}, \quad (59)$$

and the statement follows by similar arguments as above. \square

To conclude this section, we cite Lemma 42 from [23], illustrating some consequences of Corollary 6.2 and Proposition 1.3:

Example 6.3. Consider the positive definite kernels $k_{p, \sigma} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, defined via

$$k_{p, \sigma}(x, y) = \exp\left(-\frac{|x - y|^p}{\sigma^p}\right), \quad (60)$$

where $p \in (0, 2]$ is a smoothness parameter, and $\sigma > 0$ controls the kernel width. Then, following [23, Lemma 42], $k_{p, \sigma}$ is integrally strictly positive definite (see Assumption 4). Furthermore, the associated RKHSs are nested according to the regularity of the corresponding kernels: If $p > q$, then $\mathcal{H}_{k_{p, \sigma_p}} \subset \mathcal{H}_{k_{q, \sigma_q}}$ (with strict inclusion), for all $\sigma_p, \sigma_q > 0$. The inclusion of unit balls, that is,

$$\|\phi\|_{\mathcal{H}_{k_{q, \sigma_q}}} \leq \|\phi\|_{\mathcal{H}_{k_{p, \sigma_p}}}, \quad \phi \in \mathcal{H}_{k_{p, \sigma_p}}, \quad (61)$$

relevant for Proposition 1.3 can moreover be obtained by a suitable choice of the kernel widths σ_q and σ_p . Consequently, combining Proposition 1.3 and Corollary 6.2, kernels with lower regularity are expected to incur faster convergence of the ergodic limit (3) for the SDE system (6), asymptotically in the regime when N and T are large. The performance of numerical algorithms based on different choices of k is not straightforward, as the stiffness of the SDE (and corresponding time discretisations) have to be taken into account. To illustrate our findings, we instead consider fixed points of the ODE system (4) obtained for $t \rightarrow \infty$, see Figure 1. The approximation of the target π obtained using the low-regularity Laplace kernel ($p = 1$) appears to be more regular and more evenly spaced in comparison with the approximation obtained using the high-regularity squared exponential kernel ($p = 2$).

On a heuristic level, we can connect these observations to our results as follows: The large-deviation functionals $\mathcal{I}_{[0, T]}$ and $\tilde{\mathcal{I}}_T$ quantify the speed of convergence as solutions of the SDE system (6) approach solutions of the Stein PDE (8) as $N \rightarrow \infty$. Recall from Section 1.2 that the SDE (6) preserves the extended target $\tilde{\pi}$ for any $N \in \mathbb{N}$, and that solutions of the ODE system (4) solve the Stein PDE (8) in a weak sense. Therefore, our results suggests that the ODE (4) provides approximations of the SDE (6) (and hence, the target π) that are expected to be more satisfactory if $\mathcal{I}_{[0, T]}$ and $\tilde{\mathcal{I}}_T$ are large. We stress that this line of argument is heuristic and should be treated as a conjecture, since our rigorous results concern the SDE (6) and not the ODE (4). Understanding the finite-particle regime of the ODE (4), and possible connections to large-deviation principle remains an interesting subject for future research.

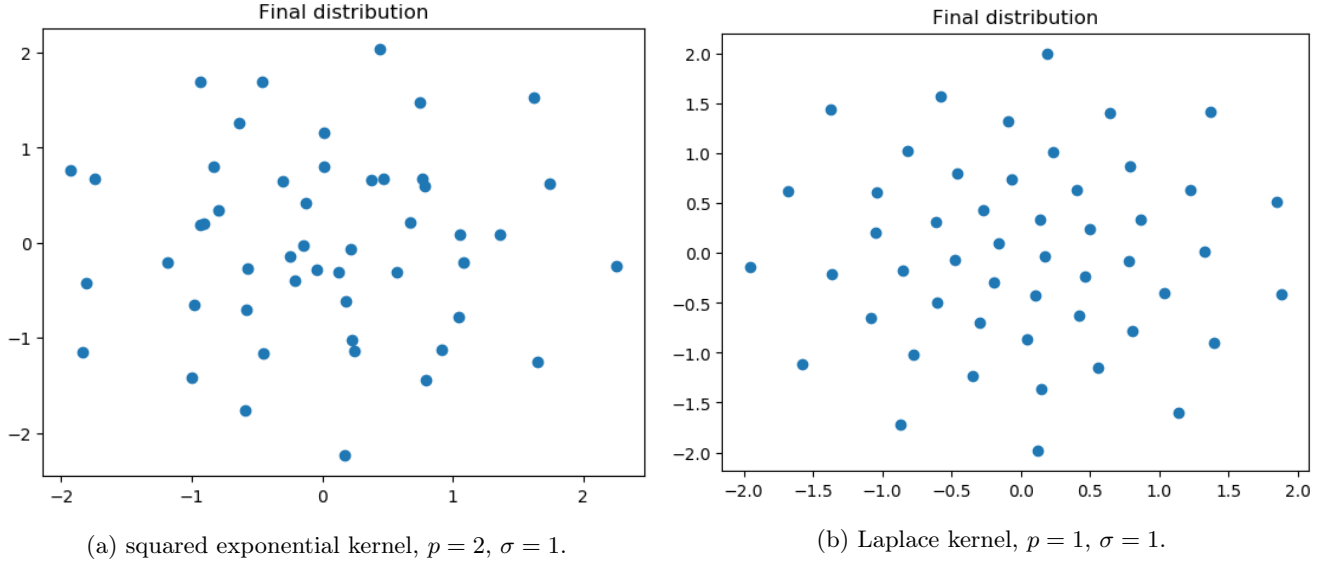


Figure 1: Approximations of a two-dimensional standard normal distribution using deterministic SVGD based on the ODE (4) and two different positive definite kernels $k_{p,\sigma}$.

7 Conclusion and outlook

In this paper, we have drawn connections between the *variational inference*-type ODE (4) and the *Markov Chain Monte Carlo*-type SDE (6) based on gradient flow structures and large-deviation functionals. Extending previous works, our results take a step towards a quantitative understanding of the mean-field limit of SVGD. In particular, in the regime when N and t are large, the convergence towards the target π is governed by the Stein-Fisher information (or kernelised Stein discrepancy). The relationship between variational inference, Markov Chain Monte Carlo and ideas from statistical physics promises to be a fruitful direction for future research beyond SVGD. As our results are asymptotic, quantifying the accuracy of SVGD for the practically relevant scenario of small N and t remains a challenging and open problem.

Acknowledgements

This research has been funded by Deutsche Forschungsgemeinschaft (DFG) through the grant CRC 1114 ‘Scaling Cascades in Complex Systems’ (projects A02 and C08, project number 235221301).

A Proofs for Section 4

Proof of Lemma 4.4. We shall only prove the claim for a large class of test functions of the form:

$$F(\rho) = \phi(\langle p_1, \rho \rangle, \dots, \langle p_L, \rho \rangle), \quad (62)$$

for arbitrary $L \in \mathbb{N}$, $p_1, \dots, p_L \in C_b^2(\mathbb{R}^d)$ and $\phi \in C_b^2(\mathbb{R}^L)$, where $\langle p_i, \rho \rangle = \int_{\mathbb{R}^d} p_i d\rho$. Applied to the empirical measure (5), these test functions become:

$$F(\rho_t^N) = \phi\left(\frac{1}{N} \sum_{i=1}^N p_1(X_t^i), \dots, \frac{1}{N} \sum_{i=1}^N p_L(X_t^i)\right) =: G(X_t^1, \dots, X_t^N) =: G(\bar{X}_t).$$

A straightforward application of Itô's Lemma to the process (6) gives, abbreviating the martingale $dM_t := \nabla G(\bar{X}_t)^T \sqrt{2K(\bar{X}_t)} dW_t$,

$$\begin{aligned} dG(\bar{X}_t) &= \sum_{i=1}^N \nabla_{X_t^i} G(\bar{X}_t) \cdot A_i(\bar{X}_t) dt + \frac{1}{N} \sum_{i,j=1}^N k(X_t^i, X_t^j) \nabla_{ij}^2 G(\bar{X}_t) dt + dM_t \\ &= \frac{1}{N} \sum_{i=1}^N \sum_{l=1}^L \partial_l \phi \nabla p_l(X_t^i) \cdot A_i(\bar{X}_t) dt + \frac{1}{N^3} \sum_{i,j=1}^N \sum_{l,m=1}^L \partial_{lm}^2 \phi \nabla p_l(X_t^i) \cdot \nabla p_m(X_t^j) k(X_t^i, X_t^j) dt \\ &\quad + \frac{1}{N^2} \sum_{l=1}^L \sum_{i=1}^N \partial_l \phi \Delta p_l(X_t^i) k(X_t^i, X_t^i) dt + dM_t, \end{aligned}$$

denoting

$$A_i(\bar{x}) = \frac{1}{N} \sum_{j=1}^N (-k(x_i, x_j) \nabla V(x_j) + \nabla_{x_j} k(x_i, x_j)), \quad \bar{x} = (x_1, \dots, x_N).$$

Notice that

$$\frac{\delta F}{\delta \rho}(\rho)(x) = \sum_{l=1}^L \partial_l \phi p_l(x), \quad \frac{\delta^2 F}{\delta \rho^2}(\rho)(x, y) = \sum_{l,m=1}^L \partial_{lm}^2 \phi p_l(x) p_m(y).$$

By taking the expectation, the martingale term drops out, so that

$$\begin{aligned} \frac{d}{dt} \mathbb{E} F(\rho_t^N) &= \mathbb{E} \left[\iint [-k(x, y) \nabla V(y) + \nabla_y k(x, y)] \cdot \nabla_x \left(\frac{\delta F}{\delta \rho}(\rho_t^N) \right) (x) \rho_t^N(dx) \rho_t^N(dy) \right] \\ &\quad + \frac{1}{N} \mathbb{E} \left[\iint k(x, y) \nabla_x \cdot \nabla_y \left(\frac{\delta^2 F}{\delta \rho^2}(\rho_t^N) \right) (x, y) \rho_t^N(dx) \rho_t^N(dy) \right. \\ &\quad \left. + \int k(x, x) \Delta \left(\frac{\delta F}{\delta \rho}(\rho_t^N) \right) (x) \rho_t^N(dx) \right], \end{aligned}$$

which proves the claim (for test functions of the form (62)). \square

Proof of Lemma 4.7. First notice that

$$\begin{aligned} \frac{\delta}{\delta \rho} (e^{NG}) (\rho)(x) &= N e^{NG(\rho)} \frac{\delta G}{\delta \rho}(\rho)(x), \quad \text{and} \\ \frac{\delta^2}{\delta \rho^2} (e^{NG}) (\rho)(x, y) &= N^2 e^{NG(\rho)} \frac{\delta G}{\delta \rho}(\rho)(x) \frac{\delta G}{\delta \rho}(\rho)(y) + N e^{NG(\rho)} \frac{\delta^2 G}{\delta \rho^2}(\rho)(x, y). \end{aligned}$$

Therefore

$$\begin{aligned} (\mathcal{H}^{(N)} G)(\rho) &\stackrel{(46)}{=} \iint_{\mathbb{R}^d \times \mathbb{R}^d} [-k(x, y) \nabla V(y) + \nabla_y k(x, y)] \cdot \left(\nabla_x \frac{\delta G}{\delta \rho}(\rho)(x) \right) \rho(dx) \rho(dy) \\ &\quad + \iint_{\mathbb{R}^d \times \mathbb{R}^d} k(x, y) \left(\nabla_x \frac{\delta G}{\delta \rho}(\rho)(x) \right) \cdot \left(\nabla_y \frac{\delta G}{\delta \rho}(\rho)(y) \right) \rho(dx) \rho(dy) + \mathcal{O}(N^{-1}). \end{aligned}$$

Assuming that ρ is regular enough, we can write

$$\int_{\mathbb{R}^d} [-k(x, y) \nabla V(y) + \nabla_y k(x, y)] \rho(dy) = - \int_{\mathbb{R}^d} k(x, y) [\nabla V(y) + \nabla \log \rho(y)] \rho(dy) = - \int_{\mathbb{R}^d} k(x, y) \frac{\delta \text{KL}}{\delta \rho}(\rho)(y) \rho(dy).$$

Then we see that

$$\begin{aligned} \mathcal{H} \left(\rho, \frac{\delta G}{\delta \rho} \right) &:= \lim_{N \rightarrow \infty} (\mathcal{H}^{(N)} G)(\rho) \\ &= - \iint_{\mathbb{R}^d \times \mathbb{R}^d} k(x, y) \left[\left(\nabla \frac{\delta \text{KL}}{\delta \rho}(\rho)(x) \right) \cdot \left(\nabla \frac{\delta G}{\delta \rho}(\rho)(y) \right) + \left(\nabla \frac{\delta G}{\delta \rho}(\rho)(x) \right) \cdot \left(\nabla \frac{\delta G}{\delta \rho}(\rho)(y) \right) \right] \rho(dx) \rho(dy) \\ &= - \left\langle \frac{\delta \text{KL}}{\delta \rho}, \frac{\delta G}{\delta \rho} \right\rangle_{T_\rho^* M} + \left\| \frac{\delta F}{\delta \rho} \right\|_{T_\rho^* M}^2. \end{aligned}$$

\square

References

- [1] L. Ambrogioni, U. Guclu, Y. Gucluturk, and M. van Gerven. Wasserstein variational gradient descent: From semi-discrete optimal transport to ensemble variational inference. *arXiv:1811.02827*, 2018.
- [2] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.
- [3] M. Arbel, A. Korba, A. Salim, and A. Gretton. Maximum mean discrepancy gradient flow. *arXiv:1906.04370*, 2019.
- [4] A. Berline and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.
- [5] L. Bertini, A. De Sole, D. Gabrielli, G. Jona-Lasinio, and C. Landim. Large deviations of the empirical current in interacting particle systems. *Theory of Probability & Its Applications*, 51(1):2–27, 2007.
- [6] P. Billingsley. *Convergence of probability measures*. Wiley, New York, NY, USA, 2nd edition, 1999.
- [7] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.
- [8] 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.
- [9] A. Braides. *Gamma convergence for beginners*. Oxford University Press, Oxford, UK, 2002.
- [10] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.
- [11] C. Chen and R. Zhang. Particle optimization in MCMC. *arXiv:1711.10927*, 2017.
- [12] C. Chen, R. Zhang, W. Wang, B. Li, and L. Chen. A unified particle-optimization framework for scalable Bayesian sampling. *arXiv:1805.11659*, 2018.
- [13] S. Chewi, T. L. Gouic, C. Lu, T. Maunu, and P. Rigollet. SVGD as a kernelized wasserstein gradient flow of the chi-squared divergence. *arXiv preprint arXiv:2006.02509*, 2020.
- [14] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *International conference on machine learning*, pages 2606–2615. PMLR, 2016.
- [15] S. Daneri and G. Savaré. Eulerian calculus for the displacement convexity in the Wasserstein distance. *SIAM Journal on Mathematical Analysis*, 40(3):1104–1122, 2008.
- [16] D. Dawson. Measure-valued Markov processes. In *Ecole d’Eté de Probabilités de Saint-Flour XXI - 1991*, pages 1–260, Berlin-Heidelberg, Germany, 1993. Springer.
- [17] D. Dawson and J. Gärtner. Large deviations from the McKean-Vlasov limit for weakly interacting diffusions. *Stochastics*, 20(4):247–308, 1987.
- [18] P. Del Moral. Feynman-Kac formulae. In *Feynman-Kac Formulae*, pages 47–93. Springer, 2004.
- [19] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38 of *Stochastic modelling and applied probability*. Springer, New York, NY, USA, 2nd edition, 1987.
- [20] H. Dietert et al. Characterisation of gradient flows on finite state Markov chains. *Electronic Communications in Probability*, 20, 2015.
- [21] A. Doucet, N. De Freitas, and N. Gordon. An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001.
- [22] J. J. Duistermaat and J. A. Kolk. *Distributions. cornerstones*, 2010.
- [23] A. Duncan, N. Nuesken, and L. Szpruch. On the geometry of Stein variational gradient descent. *arXiv preprint arXiv:1912.00894*, 2019.

- [24] J. Feng and T. Kurtz. *Large deviations for stochastic processes*, volume 131 of *Mathematical surveys and monographs*. American Mathematical Society, Providence, RI, USA, 2006.
- [25] M. A. Fisher, T. Nolan, M. M. Graham, D. Prangle, and C. J. Oates. Measure transport with kernel Stein discrepancy. *arXiv preprint arXiv:2010.11779*, 2020.
- [26] M. I. Freidlin and A. D. Wentzell. *Random perturbations of dynamical systems*, volume 260. Springer, 2012.
- [27] K. Fukumizu, A. Gretton, G. R. Lanckriet, B. Schölkopf, and B. K. Sriperumbudur. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in neural information processing systems*, pages 1750–1758, 2009.
- [28] V. Gallego and D. R. Insua. Stochastic gradient MCMC with repulsive forces. *arXiv:1812.00071*, 2018.
- [29] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler. *arXiv preprint arXiv:1903.08866*, 2019.
- [30] A. Garbuno-Inigo, N. Nüsken, and S. Reich. Affine invariant interacting Langevin dynamics for Bayesian inference. *technical report, University of Potsdam*, 2019.
- [31] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.
- [32] N. Gigli. *Second Order Analysis on $(\mathcal{P}_2(M), W_2)$* . American Mathematical Soc., 2012.
- [33] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [34] J. Gorham and L. Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR, 2017.
- [35] C. Hartmann and C. Schütte. Efficient rare event simulation by optimal nonequilibrium forcing. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(11):P11004, 2012.
- [36] M. D. Hoffman. Learning deep latent gaussian models with Markov chain Monte Carlo. In *International conference on machine learning*, pages 1510–1519, 2017.
- [37] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.
- [38] C. Kipnis and C. Landim. *Scaling limits of interacting particle systems*. Springer, Berlin-Heidelberg, Germany, 1999.
- [39] A. Korba, A. Salim, M. Arbel, G. Luise, and A. Gretton. A non-asymptotic analysis for Stein variational gradient descent. *arXiv preprint arXiv:2006.09797*, 2020.
- [40] E. Kreyszig. *Introductory functional analysis with applications*, volume 1. wiley New York, 1978.
- [41] J. M. Lee. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006.
- [42] T. Liggett. *Interacting particle systems*. Springer, Berlin-Heidelberg, Germany, 1985.
- [43] 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, 2019.
- [44] Q. Liu. Stein variational gradient descent as gradient flow. In *Advances in neural information processing systems*, pages 3115–3123, 2017.
- [45] Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284, 2016.
- [46] Q. Liu and D. Wang. Stein variational gradient descent: a general purpose Bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pages 2378–2386, 2016.

- [47] Q. Liu and D. Wang. Stein variational gradient descent as moment matching. In *Advances in Neural Information Processing Systems*, pages 8868–8877, 2018.
- [48] A. Liutkus, U. Simsekli, S. Majewski, A. Durmus, and F.-R. Stöter. Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, pages 4104–4113. PMLR, 2019.
- [49] 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, 2019.
- [50] Y.-A. Ma, T. Chen, and E. Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pages 2899–2907, 2015.
- [51] C. J. Maddison, J. Lawson, G. Tucker, N. Heess, M. Norouzi, A. Mnih, A. Doucet, and Y. Teh. Filtering variational objectives. In *Advances in Neural Information Processing Systems*, pages 6573–6583, 2017.
- [52] A. Mielke. A gradient structure for reaction–diffusion systems and for energy-drift-diffusion systems. *Nonlinearity*, 24(4):1329, 2011.
- [53] A. Mielke, M. A. Peletier, and D. R. M. Renger. On the relation between gradient flows and the large-deviation principle, with applications to Markov chains and diffusion. *Potential Analysis*, 41(4):1293–1327, 2014.
- [54] A. Mielke, D. R. M. Renger, and M. A. Peletier. A generalization of onsager’s reciprocity relations to gradient flows with nonlinear mobility. *Journal of Non-Equilibrium Thermodynamics*, 41(2):141–149, 2016.
- [55] C. Naesseth, S. Linderman, R. Ranganath, and D. Blei. Variational sequential Monte Carlo. In *International Conference on Artificial Intelligence and Statistics*, pages 968–977. PMLR, 2018.
- [56] N. Nüsken and L. Richter. Solving high-dimensional Hamilton-Jacobi-Bellman PDEs using neural networks: perspectives from the theory of controlled diffusions and measures on path space. *arXiv preprint arXiv:2005.05409*, 2020.
- [57] L. Onsager. Reciprocal relations in irreversible processes I. *Phys. Rev.*, 37(4):405–426, 1931.
- [58] L. Onsager and S. Machlup. Fluctuations and irreversible processes. *Phys. Rev.*, 91(6):1505–1512, 1953.
- [59] F. Otto. Dynamics of labyrinthine pattern formation in magnetic fluids: A mean-field theory. *Archive for Rational Mechanics and Analysis*, 141(1):63–103, 1998.
- [60] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. 2001.
- [61] F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.
- [62] F. Otto and M. Westdickenberg. Eulerian calculus for the contraction in the Wasserstein distance. *SIAM journal on mathematical analysis*, 37(4):1227–1255, 2005.
- [63] Z. Palmowski and T. Rolski. A technique for exponential change of measure for Markov processes. *Bernoulli*, 8(6):767–785, 2002.
- [64] G. A. Pavliotis. *Stochastic processes and applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, volume 60. Springer, 2014.
- [65] M. Reed and B. Simon. *Methods of modern mathematical physics: Functional analysis*. Elsevier, 2012.
- [66] S. Reich. Data assimilation: the Schrödinger perspective. *Acta Numerica*, 28:635–711, 2019.
- [67] S. Reich and C. J. Cotter. Ensemble filter techniques for intermittent data assimilation. *Large Scale Inverse Problems. Computational Methods and Applications in the Earth Sciences*, 13:91–134, 2013.
- [68] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.
- [69] W. Rudin. *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill, 2006.

- [70] F. J. Ruiz and M. K. Titsias. A contrastive divergence for combining variational inference and MCMC. *arXiv preprint arXiv:1905.04062*, 2019.
- [71] S. Saitoh and Y. Sawano. *Theory of reproducing kernels and applications*. Springer, 2016.
- [72] T. Salimans, D. Kingma, and M. Welling. Markov chain Monte Carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pages 1218–1226, 2015.
- [73] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series, 2018.
- [74] B. Schweizer. On friedrichs inequality, helmholtz decomposition, vector potentials, and the div-curl lemma. In *Trends in Applications of Mathematics to Mechanics*, pages 65–79. Springer, 2018.
- [75] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.
- [76] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.
- [77] N. G. Trillos, D. Sanz-Alonso, et al. The bayesian update: variational formulations and gradient flows. *Bayesian Analysis*, 15(1):29–56, 2020.
- [78] B. Tzen and M. Raginsky. Theoretical guarantees for sampling and inference in generative models with latent diffusions. *arXiv preprint arXiv:1903.01608*, 2019.
- [79] A. S. Üstünel and M. Zakai. *Transformation of measure on Wiener space*. Springer Science & Business Media, 2013.
- [80] C. Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.
- [81] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.