CONVERGENCE ASPECTS OF HYBRID KERNEL SVGD

Anonymous authors
Paper under double-blind review

ABSTRACT

Stein variational gradient descent (SVGD) is a particle based approximate inference algorithm. Many variants of SVGD have been proposed in recent years, including the hybrid kernel variant (h-SVGD), which has demonstrated promising results on image classification with deep neural network ensembles. In this paper, we demonstrate the ability of h-SVGD to alleviate variance collapse, a problem that SVGD is known to suffer from. Unlike other SVGD variants that alleviate variance collapse, h-SVGD does not incur additional computational cost, nor does it require the target density to factorise. We also develop the theory of h-SVGD by demonstrating the existence of a solution to the hybrid Stein partial differential equation. We highlight a special case in which h-SVGD is a kernelised Wasserstein gradient flow on a functional other than the Kullback-Leibler divergence, which is the functional describing the SVGD gradient flow. By characterising the fixed point in this special case, we show that h-SVGD does not converge to the target distribution in the the mean field limit. Other theoretical results include a descent lemma and a large particle limit result. Despite the bias in the mean field limiting distribution, experiments demonstrate that h-SVGD remains competitive on high dimensional inference tasks whilst alleviating variance collapse.

1 INTRODUCTION

Stein variational gradient descent (SVGD) is a variational inference algorithm that generates samples from a target probability density (Liu & Wang, 2016). It has proven useful in many tasks in Bayesian inference and machine learning. SVGD evolves an interacting particle system until the particles resemble a sample from a target density. The dynamics of this system include a driving term that moves particles to regions of high probability, and a repulsive term that repels particles from one another. This repulsive term prevents particles from converging to the same mode. It has been shown that within a unit ball of a reproducing kernel Hilbert space (RKHS), the SVGD update direction optimally reduces the Kullback-Leibler (KL) divergence between the target density and the approximating density (Liu & Wang, 2016). The reproducing kernel of this RKHS appears in both the driving and repulsive terms, making the choice of kernel a key ingredient for SVGD.

SVGD is known to suffer from the curse of dimensionality through variance collapse (Wang et al., 2018; Ba et al., 2019) whereby the marginal variances of the particles underestimate the true marginal variances of the target in high dimensions. Capturing the variance of the posterior is critical for uncertainty quantification in Bayesian statistics since variance is often used as a measure of confidence in an estimate. Zhuo et al. (2018) explained that this phenomenon is due to the size of the repulsive term of the update direction scaling inversely with dimension. This enables the driving term to dominate in high dimensions, thereby forcing particles to converge to the mode(s) of the target. This insight suggests that strengthening the repulsive term in SVGD should lead to better variance estimation, an idea which we explore in Sections 3 and 4.

The theoretical properties of vanilla SVGD have been studied extensively. Liu (2017) showed that the empirical measure of the particles converges weakly to the target distribution. SVGD in the mean field regime has been described as a gradient flow on the KL divergence (Liu & Wang, 2016) and the chi-squared divergence (Chewi et al., 2020). Furthering this geometric point of view, Duncan et al. (2023) developed the Stein geometry along with its associated tangent spaces and geodesics, leading to guidelines for choosing kernels to improve convergence. The existence and uniqueness of the solution to the Stein partial differential equation (PDE) was established by Lu et al. (2019).

Various descent lemmas bounding the decrease in KL divergence at each iteration have also been established (Liu, 2017; Korba et al., 2020; Salim et al., 2022).

Many variants of SVGD have also been proposed in recent years, some offering improvements and others providing generalisations. Riemannian SVGD (Liu & Zhu, 2018) generalises SVGD by allowing for target densities on Riemannian manifolds, not just Euclidean spaces. Matrix SVGD (Wang et al., 2019) replaces the scalar valued kernel with a matrix valued kernel to incorporate preconditioning information and speed up particle exploration. Message passing SVGD (Zhuo et al., 2018) and graphical SVGD (Wang et al., 2018) focus on target densities that factorise according to a graph structure. This approach reduces variance collapse in high dimensions by converting the problem to a collection of low dimensional problems. Projected SVGD (Chen & Ghattas, 2020), sliced SVGD (Gong et al., 2021) and Grassman SVGD (Liu et al., 2022) also mitigate the issue of variance collapse by updating particles within lower dimensional subspaces, which comes at the expense of additional computation.

In this work, we study a variant called hybrid kernel Stein variational gradient descent (h-SVGD). The name comes from its use of two distinct kernels for the driving and repulsive terms with the aim of mitigating variance collapse. This variant was originally proposed by D'Angelo et al. (2021) in the context of training deep neural network ensembles by sampling from the distribution of network parameters. In that setting, two particles may parameterise networks with very similar outputs despite being far apart in the weight space. Their insight was to encourage functional diversity between networks in the ensemble by using a standard kernel in the driving term, but a functional kernel in the repulsive term. In this neural network ensemble setting, h-SVGD demonstrated better performance than other variants on image classification. Annealed SVGD (D'Angelo & Fortuin, 2021) may also be considered an example of h-SVGD. In this variant, the driving kernel is a scalar multiple $\gamma(\ell) \in [0, 1]$ of the repulsive kernel, and this factor $\gamma(\ell)$ gradually increases to 1 as the iteration ℓ increases. Numerical experiments show that annealed SVGD improves the ability of particles to escape local modes. Scaling one of the update terms has also been used as a computational technique to aid other SVGD variants when training Bayesian neural networks (Gong et al., 2021).

Although preliminary numerical experiments have shown the benefits of h-SVGD (D'Angelo et al., 2021), the theoretical results for SVGD do not directly apply in the hybrid kernel setting due to the presence of a second kernel. In this paper, we address this theoretical gap and reinforce the practical benefits of h-SVGD through the following contributions.

- We establish existence of a solution to the hybrid Stein PDE and a kernelised Wasserstein gradient flow interpretation. Through the study of this gradient flow, we demonstrate that h-SVGD does *not* converge to the target distribution in the mean field limit.
- Other theoretical contributions include quantifying the rate of dissipation of the gradient flow functional, along with a discrete time version, otherwise known as a descent lemma.
- Despite not converging to the target distribution, numerical experiments show that h-SVGD can mitigate variance collapse in the finite particle regime at negligible additional cost, whilst remaining competitive at high dimensional inference tasks.

In Section 2 we clarify notation, recall necessary theory, and outline the vanilla and hybrid SVGD algorithms. Section 3 contains the theoretical contributions, with proofs relegated to Appendix A. Numerical experiments are in Section 4 with additional experiments and details in Appendix B.

2 Background

2.1 Notation

Let $\mathcal{X} \subseteq \mathbb{R}^d$. Let π denote the target probability density on $\mathcal{X}$ and let $\boldsymbol{s}_\pi(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log \pi(\boldsymbol{x})$. We will often write π for the corresponding measure as well. Assume that $\pi(\boldsymbol{x}) = e^{-V(\boldsymbol{x})}$ for some potential V. Let $\mathcal{P}(\mathcal{X})$ be the set of probability measures on $\mathcal{X}$ and let $\mathcal{P}_V(\mathcal{X})$ denote the subset where $\|\mu\|_{\mathcal{P}_V} := \int_{\mathcal{X}} (1 + V(\boldsymbol{x})) d\mu(\boldsymbol{x}) < \infty$. For $p \geq 1$, let $\mathcal{P}_p(\mathcal{X})$ be the subset where $\|\mu\|_{\mathcal{P}_p} := \int_{\mathcal{X}} \|\boldsymbol{x}\|^p d\mu(\boldsymbol{x}) < \infty$ and define the Wasserstein p-distance between the two measures $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ as $W_p(\mu, \nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int \|\boldsymbol{x} - \boldsymbol{y}\|^p d\gamma(\mu, \nu)\right)^{1/p}$, where $\Gamma(\mu, \nu)$ is the set of couplings between μ and ν. Let $L^\infty_c(\mathcal{X})$ denote the set of probability densities bounded almost

everywhere with compact support. Let $L^2(\mu)$ denote the set of functions that are square integrable with respect to the measure μ. Given $\mu \in \mathcal{P}(\mathcal{X})$ and a smooth, invertible transform $\boldsymbol{T} : \mathcal{X} \to \mathcal{X}$, let $\boldsymbol{T}_{\#}\mu$ denote the pushforward measure of μ through $\boldsymbol{T}$. The KL divergence between two measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is denoted by $\mathrm{KL}(\mu \parallel \nu)$. Fourier transforms are denoted by $\mathscr{F}$.

2.2 Reproducing Kernel Hilbert Spaces

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if $\sum_{i,j} a_i k(\boldsymbol{x}_i, \boldsymbol{x}_j) a_j > 0$ for any $a_1, \ldots, a_d \in \mathbb{R}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d \in \mathcal{X}$. Given a Hilbert space $\mathcal{H}$ of functions $\phi : \mathcal{X} \to \mathbb{R}$, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a reproducing kernel for $\mathcal{H}$ if it satisfies the reproducing property, $\phi(\boldsymbol{x}) = \langle \phi, k(\boldsymbol{x}, \cdot) \rangle_{\mathcal{H}}$ for all $\phi \in \mathcal{H}$. A positive definite $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ admits a unique Hilbert space $\mathcal{H}$ of functions $\phi : \mathcal{X} \to \mathbb{R}$ for which the Dirac functionals $\delta_{\boldsymbol{x}} : \mathcal{H} \to \mathbb{R}, \delta_{\boldsymbol{x}} \phi = \phi(\boldsymbol{x})$ are all continuous and k is a reproducing kernel. This Hilbert space is called the reproducing kernel Hilbert space (RKHS) of k and it is equal to the closure of the span of $\{k(\boldsymbol{x}, \cdot) : \boldsymbol{x} \in \mathbb{R}\}$. Let $\mathcal{H}^d = \mathcal{H} \times \cdots \times \mathcal{H}$ denote the Hilbert space of functions $\phi : \mathcal{X} \to \mathbb{R}^d$ whose components are all in $\mathcal{H}$, and equip it with the usual inner product $\langle \phi, \psi \rangle_{\mathcal{H}^d} = \sum_{i=1}^d \langle \phi_i, \psi_i \rangle_{\mathcal{H}}$. Given two kernels $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, let $\mathcal{H}_1, \mathcal{H}_2$ denote their respective RKHS. An important kernel used throughout this paper is the radial basis function (RBF) kernel $k_{\mathrm{RBF}}(\boldsymbol{x}, \boldsymbol{y}; h) := \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 / (2h))$ with bandwidth $h > 0$. For a thorough treatment of RKHS we refer the reader to (Aronszajn, 1950; Steinwart & Christmann, 2008; Berlinet & Thomas-Agnan, 2011).

2.3 Stein Variational Gradient Descent

The key result from (Liu & Wang, 2016) identifies a transform $\boldsymbol{T} : \mathcal{X} \to \mathcal{X}$ that optimally decreases the KL divergence from an arbitrary probability measure to π. More precisely, let $\mathcal{H}$ be an RKHS with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and consider transforms of the form $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{x} + \epsilon \phi(\boldsymbol{x})$ where $\epsilon > 0$ and ϕ is in the unit ball $\{\phi \in \mathcal{H}^d : \|\phi\|_{\mathcal{H}^d} \leq 1\}$. The maximum value of

$$-\nabla_\epsilon \mathrm{KL}(\boldsymbol{T}_{\#}\mu \parallel \pi)|_{\epsilon=0} \tag{1}$$

occurs at $\phi^k_{\mu,\pi} / \left\|\phi^k_{\mu,\pi}\right\|_{\mathcal{H}^d}$, where

$$\phi^k_{\mu,\pi}(\cdot) := \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[k(\boldsymbol{x}, \cdot)\boldsymbol{s}_\pi(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \cdot)\right]. \tag{2}$$

When μ is an empirical distribution (i.e. a sum of Dirac measures), the expectation in (2) can be computed exactly by summing over the particles of each Dirac measure. Using this observation, the SVGD algorithm starts with an initial set of N particles $(\boldsymbol{x}^i_0)_{i=1}^N$ and iteratively applies the transform $\boldsymbol{T}$ with (2) as the update direction. At each iteration ℓ, this yields a set of particles $(\boldsymbol{x}^i_\ell)_{i=1}^N$ and a corresponding empirical distribution $\mu_\ell = \frac{1}{N}\sum_i \delta_{\boldsymbol{x}^i_\ell}$. This is captured in Algorithm 1. The intention is that after sufficiently many iterations, the set of particles will resemble samples from π and expectations of the form $\mathbb{E}_{\boldsymbol{x}\sim\pi} h(\boldsymbol{x})$ can be approximated by $\mathbb{E}_{\boldsymbol{x}\sim\mu_\ell} h(\boldsymbol{x}) = \frac{1}{N}\sum_i h(\boldsymbol{x}^i_\ell)$. We also recall the definition of the kernelised Stein discrepancy (KSD) from (Liu et al., 2016),

$$\mathbb{S}_k(\mu, \pi) := \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\mu}\left[(\boldsymbol{s}_\pi(\boldsymbol{x}) - \boldsymbol{s}_\mu(\boldsymbol{x}))^\top k(\boldsymbol{x}, \boldsymbol{y})(\boldsymbol{s}_\pi(\boldsymbol{y}) - \boldsymbol{s}_\mu(\boldsymbol{y}))\right]. \tag{3}$$

Algorithm 1 Stein Variational Gradient Descent (Liu & Wang, 2016)

Input: A target probability distribution π, a kernel k, an initial set of particles $(\boldsymbol{x}^i_0)_{i=1}^N$ in $\mathcal{X}$, and a sequence of step sizes (ϵ_ℓ).
Output: A set of particles $(\boldsymbol{x}^i)_{i=1}^N$ in $\mathcal{X}$ whose empirical distribution approximates π.
for iteration ℓ **do**

$$\boldsymbol{x}^i_{\ell+1} \leftarrow \boldsymbol{x}^i_\ell + \epsilon_\ell \hat{\phi}^*_{\mu_\ell,\pi}(\boldsymbol{x}^i_\ell), \qquad \forall\ i = 1, \ldots, N$$

$$\hat{\phi}^*_{\mu_\ell,\pi}(\boldsymbol{x}) = \frac{1}{N}\sum_{j=1}^N \left(k(\boldsymbol{x}^j_\ell, \boldsymbol{x})\nabla_{\boldsymbol{x}^j_\ell} \log \pi(\boldsymbol{x}^j_\ell) + \nabla_{\boldsymbol{x}^j_\ell} k(\boldsymbol{x}^j_\ell, \boldsymbol{x})\right) \tag{4}$$

2.4 HYBRID KERNEL STEIN VARIATIONAL GRADIENT DESCENT

The SVGD update in (4) contains two terms, each using the same kernel. The first term, often referred to as the driving term, uses the score function to move particles towards regions of high probability density, and the repulsive term prevents particles from collapsing at the modes. The h-SVGD variant proposed by D'Angelo et al. (2021) uses a different kernel in each term. Let k_1 denote the kernel that appears alongside the score function, and let k_2 denote the repulsive kernel. For the remainder of this paper, k_1 and k_2 will both be positive definite. We present h-SVGD in Algorithm 2.

Algorithm 2 Hybrid Kernel Stein Variational Gradient Descent

Input: A target probability distribution π, two kernels k_1, k_2, an initial set of particles $(\boldsymbol{x}_0^i)_{i=1}^N$ in $\mathcal{X}$, and a sequence of step sizes (ϵ_ℓ).
Output: A set of particles $(\boldsymbol{x}^i)_{i=1}^N$ in $\mathcal{X}$ whose empirical distribution approximates π.
for iteration ℓ **do**

$$\boldsymbol{x}_{\ell+1}^i \leftarrow \boldsymbol{x}_\ell^i + \epsilon_\ell \hat{\phi}^*_{\mu_\ell,\pi}(\boldsymbol{x}_\ell^i), \qquad \forall i = 1, \dots, N$$

$$\hat{\phi}^*_{\mu_\ell,\pi}(\boldsymbol{x}) = \frac{1}{N}\sum_{j=1}^{N}\left(k_1(\boldsymbol{x}_\ell^j, \boldsymbol{x})\nabla_{\boldsymbol{x}_\ell^j}\log\pi(\boldsymbol{x}_\ell^j) + \nabla_{\boldsymbol{x}_\ell^j}k_2(\boldsymbol{x}_\ell^j, \boldsymbol{x})\right) \tag{5}$$

3 THEORETICAL RESULTS

3.1 DEFINITIONS AND ASSUMPTIONS

A function $f : \mathcal{X} \to \mathbb{R}$ is in the Stein class of π if it is smooth and satisfies the identity $\int_{\boldsymbol{x}\in\mathcal{X}} \nabla_{\boldsymbol{x}} (f(\boldsymbol{x})\pi(\boldsymbol{x}))\, d\boldsymbol{x} = 0$. A function $\boldsymbol{f} = (f_1, \dots, f_d) : \mathcal{X} \to \mathbb{R}^d$ is in the Stein class of π if each f_i is in the Stein class of π. A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is in the Stein class of π if it has continuous second order partial derivatives and both $k(\boldsymbol{x}, \cdot)$ and $k(\cdot, \boldsymbol{y})$ are in the Stein class of π for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$. The hybrid Stein operator acts on a pair of kernels $k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$\mathcal{S}_\pi \otimes (k_1, k_2)(\boldsymbol{x}, \cdot) := k_1(\boldsymbol{x}, \cdot)\boldsymbol{s}_\pi(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} k_2(\boldsymbol{x}, \cdot),$$

provided k_1 and k_2 both belong to the Stein class of π. This reduces to the Stein operator defined in (Liu et al., 2016) when $k_1 = k_2$. Motivated by the h-SVGD update in (5), define the update direction

$$\phi^{k_1,k_2}_{\mu,\pi}(\cdot) := \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\mathcal{S}_\pi \otimes (k_1, k_2)(\boldsymbol{x}, \cdot)\right], \tag{6}$$

and write $\phi^* = \phi^{k_1,k_2}_{\mu,\pi} / \left\|\phi^{k_1,k_2}_{\mu,\pi}(\cdot)\right\|_{\mathcal{H}_1}$. Let $\boldsymbol{G}(\,\cdot\,; k_1, \mu, \pi) := \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[k_1(\boldsymbol{x}, \cdot)\boldsymbol{s}_\pi(\boldsymbol{x})\right]$ denote the gradient term and $\boldsymbol{R}(\,\cdot\,; k_2, \mu) := \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\nabla_{\boldsymbol{x}} k_2(\boldsymbol{x}, \cdot)\right]$ the repulsive term. We can then write $\phi^{k_1,k_2}_{\mu,\pi}(\cdot) = \boldsymbol{G}(\,\cdot\,; k_1, \mu, \pi) + \boldsymbol{R}(\,\cdot\,; k_2, \mu)$. The update transform

$$\boldsymbol{T}^{k_1,k_2}_{\mu,\pi}(\boldsymbol{x}) = \boldsymbol{x} + \epsilon\phi^{k_1,k_2}_{\mu,\pi}(\boldsymbol{x}) \tag{7}$$

and the map $\Phi^{k_1,k_2}_\pi : \mu \mapsto \left(\boldsymbol{T}^{k_1,k_2}_{\mu,\pi}\right)_{\#}\mu$ characterise the h-SVGD dynamics. For each ℓ, define

$$\mu^N_{\ell+1} := \Phi^{k_1,k_2}_\pi(\mu^N_\ell), \qquad \mu^\infty_{\ell+1} := \Phi^{k_1,k_2}_\pi(\mu^\infty_\ell), \tag{8}$$

where μ_0^N is the empirical measure of the initial particles $(\boldsymbol{x}_0^i)_{i=1}^N$ drawn i.i.d. from some μ_0^∞.

All technical assumptions required in the theorems throughout this section are detailed here for completeness. The first set of assumptions relate to the potential of the target distribution.

(**A**1) $V \in C^\infty(\mathcal{X})$, $V \geq 0$, and $\lim_{|\boldsymbol{x}|\to\infty} V(\boldsymbol{x}) = +\infty$.

(**A**2) There exist constants $C_V > 0$ and $q > 1$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,

$$|\nabla V(\boldsymbol{x})|^q \leq C_V(1 + V(\boldsymbol{x}))$$

and

$$\sup_{\theta\in[0,1]} \left|\nabla^2 V(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y})\right|^q \leq C_V(1 + V(\boldsymbol{x}) + V(\boldsymbol{y})).$$

(**A**3) For any $\alpha, \beta > 0$, there exists a constant $C_{\alpha,\beta} > 0$ such that

$$|\boldsymbol{y}| \leq \alpha\,|\boldsymbol{x}| + \beta \implies (1 + |\boldsymbol{x}|)(|\nabla V(\boldsymbol{y})| + \left|\nabla^2 V(\boldsymbol{y})\right|) \leq C_{\alpha,\beta}(1 + V(\boldsymbol{x})).$$

(**A**4) The Hessian H_V of V is well-defined and $\|H_V\|_{\mathrm{op}} \leq M$ for some $M > 0$.

Assumptions (**A**1), (**A**2) and (**A**3) are identical to those in (Lu et al., 2019). Assumption (**A**4) is identical to Assumption (A2) in (Korba et al., 2020), and Assumption 2.1 in (Salim et al., 2022). We remark that Assumptions (**A**2) and (**A**3) simply control the decay of the tails of π. The above assumptions are all satisfied by normal distributions, mixtures thereof, as well as many continuous distributions from the exponential family. The following assumptions are also required to ensure sufficient regularity of the kernel functions.

(**B**1) There exist symmetric functions $K_1, K_2 : \mathcal{X} \to \mathbb{R}$ such that $k_1(\boldsymbol{x}, \boldsymbol{y}) = K_1(\boldsymbol{x} - \boldsymbol{y})$, $k_2(\boldsymbol{x}, \boldsymbol{y}) = K_2(\boldsymbol{x} - \boldsymbol{y})$, K_1 is C^2 with bounded derivatives, and K_2 is C^4 with bounded derivatives. By symmetric, we mean $K_i(\boldsymbol{x}) = K_i(-\boldsymbol{x})$ for all $i = 1, 2$ and $\boldsymbol{x} \in \mathcal{X}$. We use $B > 0$ as a bound for all derivatives in the proofs.

(**B**2) There exists a constant $D > 0$ such that both k_1 and ∇k_2 and are D-Lipschitz, and $\nabla V(\cdot)k_1(\cdot, \boldsymbol{z})$ is D-Lipschitz for each $\boldsymbol{z}$. That is,

$$\begin{aligned} |k_1(\boldsymbol{x}, \boldsymbol{x}') - k_1(\boldsymbol{y}, \boldsymbol{y}')| &\leq D\left(\|\boldsymbol{x} - \boldsymbol{y}\|_2 + \|\boldsymbol{x}' - \boldsymbol{y}'\|_2\right), \\ \|\nabla_{\boldsymbol{x}} k_2(\boldsymbol{x}, \boldsymbol{x}') - \nabla_{\boldsymbol{y}} k_2(\boldsymbol{y}, \boldsymbol{y}')\| &\leq D\left(\|\boldsymbol{x} - \boldsymbol{y}\|_2 + \|\boldsymbol{x}' - \boldsymbol{y}'\|_2\right), \\ \|\nabla V(\boldsymbol{x})k_1(\boldsymbol{x}, \boldsymbol{z}) - \nabla V(\boldsymbol{y})k_1(\boldsymbol{y}, \boldsymbol{z})\| &\leq D\left(\|\boldsymbol{x} - \boldsymbol{y}\|_2\right) \end{aligned}$$

for all $\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{y}, \boldsymbol{y}', \boldsymbol{z} \in \mathcal{X}$.

Assumption (**B**1) is a slight relaxation of Assumption 2.1 in (Lu et al., 2019), and contains Assumption 2.6 in (Salim et al., 2022). The first two parts of Assumption (**B**2) are hybrid kernel versions of Assumption (B2) from (Korba et al., 2020), and the third part of Assumption (**B**2) relaxes the restrictive Assumption (B1) from (Korba et al., 2020).

3.2 Large Particle Limit

We begin our theoretical study with a result ensuring convergence of h-SVGD in the large particle limit. The single kernel version (Korba et al., 2020, Proposition 7) of the following result uses an assumption that is quite restrictive. It requires $|V(\boldsymbol{x})| \leq C_V$ for some constant $C_V > 0$, which rules out even a normal target distribution. We relax this with the third part of Assumption (**B**2) and provide an updated proof in the appendix.

Proposition 3.1. *Assume (**A**1), (**A**4), (**B**1) and (**B**2), and let $T > 0$. For any $0 \leq \ell \leq \frac{T}{\epsilon_\ell}$, there exists a constant L depending on k_1, k_2 and π such that*

$$\mathbb{E}\left[W_2^2(\mu_\ell^N, \mu_\ell^\infty)\right] \leq \frac{1}{2\sqrt{N}}\sqrt{\mathrm{var}(\mu_0^\infty)}e^{LT}(e^{2LT} - 1).$$

3.3 Hybrid Stein PDE

In the continuous time limit, equation (5) becomes a coupled system of differential equations,

$$\frac{d\boldsymbol{x}_i}{dt} = \frac{1}{N}\sum_{j=1}^{N}\left(k_1(\boldsymbol{x}_j, \boldsymbol{x}_i)\nabla_{\boldsymbol{x}_j}\log\pi(\boldsymbol{x}_j) + \nabla_{\boldsymbol{x}_j}k_2(\boldsymbol{x}_j, \boldsymbol{x}_i)\right), \qquad i = 1, \ldots, N. \tag{9}$$

In the mean field limit, integration by parts yields

$$\begin{aligned} \frac{d\boldsymbol{x}}{dt} &= \int\left(k_1(\boldsymbol{x}', \boldsymbol{x})\nabla_{\boldsymbol{x}'}\log\pi(\boldsymbol{x}') + \nabla_{\boldsymbol{x}'}k_2(\boldsymbol{x}', \boldsymbol{x})\right)\rho(\boldsymbol{x}')d\boldsymbol{x}' &(10)\\ &= T_{k_1,\rho}\left(\nabla\log\pi\right)(\boldsymbol{x}) - T_{k_2,\rho}\left(\nabla\log\rho\right)(\boldsymbol{x}) &(11)\end{aligned}$$

where $T_{k,\rho} : L^2(\rho)^d \to \mathcal{H}_k^d$ is the Hilbert-Schmidt operator $(T_{k,\rho}f)(\cdot) = \int k(\cdot, \boldsymbol{x})f(\boldsymbol{x})d\rho(\boldsymbol{x})$ for a kernel k. Recalling that V is the potential of π and so $V = -\log\pi$, this mean field limit can be described by the hybrid Stein partial differential equation

$$\partial_t\rho_t = \nabla\cdot\left(\rho_t\left(K_1 * \nabla V\rho_t + K_2 * \nabla\rho_t\right)\right). \tag{12}$$

Definition 3.1. *Given a probability measure ν on $\mathbb{R}^d$, a map $X(t, \boldsymbol{x}; \nu) : [0, \infty) \times \mathbb{R}^d \to \mathbb{R}^d$ that is C^1 with respect to t and satisfies*

$$\begin{aligned} \partial_t X(t, \boldsymbol{x}; \nu) &= -\left(K_1 * (\nabla V \rho_t)\right)\left(X(t, \boldsymbol{x}; \nu)\right) - \left(\nabla K_2 * \rho_t\right)\left(X(t, \boldsymbol{x}; \nu)\right) \\ \rho_t &= X(t, \cdot, \nu)_{\#}\nu \\ X(0, \boldsymbol{x}; \nu) &= \boldsymbol{x} \end{aligned} \tag{13}$$

is called a mean field characteristic flow of (10) or of (12).

We now generalise Theorem 2.4 from (Lu et al., 2019), ensuring the existence of a solution to the hybrid Stein PDE. First, define the set $Y := \{u \in C(\mathcal{X}, \mathcal{X}) : \sup_{\boldsymbol{x} \in \mathcal{X}} |u(\boldsymbol{x}) - \boldsymbol{x}| < \infty\}$ with $d_Y(u, v) = \sup_{\boldsymbol{x} \in \mathcal{X}} |u(\boldsymbol{x}) - v(\boldsymbol{x})|$ and note that (Y, d_Y) is a complete metric space.

Proposition 3.2. *Assume (**A**1), (**A**2), (**A**3) and (**B**1), and let $T > 0$. Then there is a unique solution $X(\cdot, \cdot, \nu) \in C^1([0, T]; Y)$ to (13) and the corresponding ρ_t is a weak solution to (12) satisfying*

$$\|\rho_t\|_{\mathcal{P}_V} \leq \|\pi\|_{\mathcal{P}_V} \exp\left(C \min\left(\|\nabla K_1\|_\infty, \|\nabla K_2\|_\infty\right) t\right) \tag{14}$$

for some constant $C > 0$ depending on K_1, K_2 and V.

The second kernel enables a stronger bound than Theorem 2.4 from (Lu et al., 2019) by careful modification of the proof (see Appendix A). In particular, ensuring that $\|\nabla K_1\|_\infty < \|\nabla K_2\|_\infty$ when choosing K_2 yields a stronger bound in (14) than if K_1 were used for both kernels. We remark that this bound describes regularity of the solution to the PDE, not a rate of convergence.

3.4 Kernelised Wasserstein Gradient Flow and Asymptotic Density: $k_2 = ck_1$

Zhuo et al. (2018) uncovered a correlation between dimension and the magnitude of the repulsive force. Under some technical conditions, for any $\alpha, \delta \in (0, 1)$, they show that SVGD under an RBF kernel yields $\|\boldsymbol{R}(\,\cdot\,; k_2, \mu)\|_\infty = O(d^{-\alpha})$ with probability at least $1 - \delta$. This suggests that simply scaling the repulsive force by d^α for some $\alpha \in (0, 1)$ should offset the decrease in $\|\boldsymbol{R}(\,\cdot\,; k_2, \mu)\|_\infty$ in high dimensions, thereby alleviating variance collapse at negligible additional computational cost. Scaling the repulsive kernel in this way corresponds to h-SVGD where k_1 is an RBF kernel and $k_2 = d^\alpha k_1$. This motivates our study of the h-SVGD gradient flow under the special case $k_2 = ck_1$.

Recall that the Wasserstein gradient flow on a functional $\mathcal{F}$ is the time evolution of the probability measure ρ_t that minimises $\mathcal{F}(\rho)$. This is described by a PDE and the Wasserstein gradient is defined through the Wasserstein distance W_2. We refer the reader to (Ambrosio et al., 2005; Santambrogio, 2015) for further details.

In the case where $k = k_1 = k_2$, (10) and (12) describe a kernelised Wasserstein gradient flow of the form $\partial_t \rho_t = \nabla \cdot (\rho_t T_{k, \rho_t} \nabla_W \mathcal{F}(\rho_t))$, where

$$\mathcal{F}(\rho) = \mathrm{KL}(\rho \parallel \pi) = \mathbb{E}_{\boldsymbol{x} \sim \rho}\left[\log \rho(\boldsymbol{x}) - \log \pi(\boldsymbol{x})\right]$$

is the KL divergence functional (Liu, 2017). This has functional derivative $\frac{\delta \mathcal{F}}{\delta \rho} = \log \rho - \log p + 1$ up to a constant, and Wasserstein gradient

$$\nabla_W \mathcal{F}(\rho) = \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\rho) = \nabla \log \rho - \nabla \log \pi. \tag{15}$$

Then (11) can be written as

$$\frac{d\boldsymbol{x}}{dt} = -T_{k, \rho} \nabla_W \mathcal{F}(\rho),$$

and the corresponding Fokker-Planck equation is

$$\partial_t \rho_t = \nabla \cdot \left(\rho_t T_{k, \rho_t}\left(\nabla \log \rho_t - \nabla \log \pi\right)\right) \tag{16}$$

where $\{\rho_t : t \geq 0\}$ is a curve of probability densities. The following result generalises this gradient flow interpretation to the case where $k_2 = ck_1$. Note that it applies to any positive definite kernel k_1, not just the RBF kernel as discussed earlier.

Proposition 3.3. *Given a positive definite kernel k_1 and constant $c > 0$, let $k_2 = ck_1$. Then the mean field dynamics of h-SVGD describe a kernelised Wasserstein gradient flow on the functional*

$$\mathcal{F}(\rho) = \mathbb{E}_{\boldsymbol{x}\sim\rho}\left[c\log\rho - \log\pi(\boldsymbol{x})\right], \tag{17}$$

whose Wasserstein gradient is

$$\nabla_W \mathcal{F}(\rho) = c\nabla\log\rho - \nabla\log\pi. \tag{18}$$

The corresponding continuity equation is

$$\partial_t \rho_t = \nabla\cdot\left(\rho_t T_{k_1,\rho_t}\left(c\nabla\log\rho_t - \nabla\log\pi\right)\right). \tag{19}$$

Even in this simple hybrid kernel setting, the following result establishes that the limiting distribution ρ^* of the mean field regime is not equal to the target distribution π.

Corollary 3.4. *If $k_2 = ck_1$ for some constant $c > 0$ where k_1 is a positive definite kernel, then the mean field h-SVGD has a fixed point $\rho^*(\boldsymbol{x}) \propto \pi(\boldsymbol{x})^{1/c}$.*

Although Corollary 3.4 applies to any target density π satisfying (**A**1), it is especially insightful to consider a normal target. If $\pi(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$, then $\rho^*(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},c\boldsymbol{\Sigma})$. So scaling the repulsive kernel k_2 will adjust the variance of the target by the same factor. This supports the motivation that scaling the repulsive kernel should offset the variance underestimation in high dimensions at a negligible additional cost. This idea will be revisited in Section 4.

We now generalise an existing result (Korba et al., 2020, Proposition 1) that describes the dissipation of the KL divergence along the SVGD gradient flow. The result below describes the dissipation of the functional in (17) along the h-SVGD gradient flow, ensuring that the functional always decreases. It also describes the dissipation of the KL divergence with respect to the mean field limiting distribution ρ^*, which we emphasise is not equal to the target distribution π, as per Corollary 3.4.

Proposition 3.5. *Under the assumptions of Proposition 3.3,*

$$\frac{d}{dt}\mathcal{F}(\rho_t) = -c^2\left\|T_{k_1,\rho_t}(\nabla\log\rho - \nabla\log\rho^*)\right\|^2_{\mathcal{H}_1^d} + c\int\rho_t(\boldsymbol{x})\frac{\partial}{\partial t}\log\rho_t(\boldsymbol{x})d\boldsymbol{x}, \tag{20}$$

$$\frac{d}{dt}\mathrm{KL}(\rho_t\parallel\rho^*) = -c\left\|T_{k_1,\rho_t}(\nabla\log\rho - \nabla\log\rho^*)\right\|^2_{\mathcal{H}_1^d} + \int\rho_t(\boldsymbol{x})\frac{\partial}{\partial t}\log\rho_t(\boldsymbol{x})d\boldsymbol{x}, \tag{21}$$

where $\rho^*(\boldsymbol{x}) \propto \pi(\boldsymbol{x})^{1/c}$ is the mean field fixed point. Furthermore,

$$\int\rho_t(\boldsymbol{x})\frac{\partial}{\partial t}\log\rho_t(\boldsymbol{x})d\boldsymbol{x} \le 0,$$

so $\frac{d}{dt}\mathcal{F}(\rho_t) \le 0$ and $\frac{d}{dt}\mathrm{KL}(\rho\parallel\rho^*) \le 0$ for $t \ge 0$, and $\frac{d}{dt}\mathcal{F}(\rho^*) = 0$.

The following descent lemma is adapted from (Liu, 2017, Theorem 3.3) and it provides a discrete time version of Proposition 3.5. We use μ_ℓ to denote discrete time steps of the algorithm, as defined in (8), as opposed to ρ_t for the continuous time analysis. We remark that other descent lemmas for SVGD have been established (Korba et al., 2020; Salim et al., 2022).

Proposition 3.6. *Set $\epsilon_\ell = (2\sup_x\sigma(\nabla\phi^* + \nabla\phi^{*\top}))^{-1}$ where $\sigma(\cdot)$ denotes the spectral radius of a matrix. Define the quantity $R = \sup_{\boldsymbol{x}}\{\frac{1}{2}\left\|\nabla\log\pi\right\|_{\mathrm{Lip}}k(\boldsymbol{x},\boldsymbol{x}) + 2\nabla_{\boldsymbol{x},\boldsymbol{y}}k(\boldsymbol{x},\boldsymbol{x})\}$ where $\nabla_{\boldsymbol{x},\boldsymbol{y}}k(\boldsymbol{x},\boldsymbol{x}) = \sum_i\partial_{x_i}\partial_{y_i}k(\boldsymbol{x},\boldsymbol{y})|_{\boldsymbol{x}=\boldsymbol{y}}$. Then*

$$\frac{1}{\epsilon_\ell}\left(\mathcal{F}(\mu^\infty_{\ell+1}) - \mathcal{F}(\mu^\infty_\ell)\right) \le -(1-\epsilon_\ell R)c\mathbb{S}(\mu_\ell,\rho^*).$$

This result ensures that for a sufficiently small step size ϵ_ℓ, the functional $\mathcal{F}$ will decrease at each step of the algorithm, until μ_ℓ approaches ρ^*.

3.5 Kernelised Wasserstein Gradient Flow: The General Case

In this section, we present a generalisation of Proposition (3.3) and discuss some difficulties in finding kernels that satisfy the required conditions. For ease of presentation, we restrict our attention

to $d = 1$. We assume (**B**1) throughout, and also assume that any required Fourier transforms exist. Define the function $r : \mathcal{X} \to \mathbb{R}$ by

$$r(x; \rho) := \mathscr{F}^{-1}\left(\frac{\mathscr{F}(K_2)}{\mathscr{F}(K_1)} \cdot \mathscr{F}(\nabla\rho)\right)(x) \tag{22}$$

and let $R : \mathcal{X} \to \mathbb{R}$ denote a function satisfying $\nabla R(x; \rho) = r(x; \rho)/\rho(x)$.

Proposition 3.7. *Assume that both r and R exist. Then the corresponding continuity equation is*

$$\partial_t \rho_t = \nabla \cdot (\rho_t T_{k_1,\rho_t} (\nabla R(\cdot\,; \rho_t) - \nabla \log \pi)) . \tag{23}$$

If in addition

$$\int \frac{\partial}{\partial\epsilon} R(x; \rho + \epsilon\chi) d(\rho + \epsilon\chi)(x)\Big|_{\epsilon=0} = 0 \tag{24}$$

for any measure $\chi = \tilde{\rho} - \rho$ with $\tilde{\rho} \in L_c^\infty(\mathcal{X}) \cap \mathcal{P}(\mathcal{X})$, then the mean field dynamics of h-SVGD describe a kernelised Wasserstein gradient flow on the functional

$$\mathcal{F}(\rho) = \mathbb{E}_{x\sim\rho} [R(x; \rho) - \log \pi(x)] , \tag{25}$$

whose Wasserstein gradient is

$$\nabla_W \mathcal{F}(\rho) = r(x; \rho)/\rho(x) - \nabla \log \pi(x). \tag{26}$$

Note that when $k_2 = ck_1$, we have $\mathscr{F}(K_2)/\mathscr{F}(K_1) = c$ and so $r(x; \rho) = c\nabla\rho(x)$. Therefore, $R(x; \rho) = c \log \rho(x)$, and so (23) reduces to (19). Furthermore, the left hand side of (24) is equal to $c \int d\chi(\boldsymbol{x}) = c \int d\tilde{\rho}(x) - c \int d\rho(x) = 0$ because $\tilde{\rho}, \rho$ are both probability measures. So (24) is satisfied in this special case.

We remark that verifying (24) remains a challenge in general hybrid kernel settings. However, the derivation of the continuity equation does not rely on (24), so we can gain some insights into the steady state by studying when (23) is equal to zero. We present a specific example below to demonstrate the behaviour of SVGD in the general hybrid kernel setting. In particular, the form of the target π and the mean field steady state ρ^* can look quite different, even under a hybrid kernel setting as simple as two RBF kernels with different bandwidths.

Proposition 3.8. *Let $h_1, h_2, \sigma > 0$ such that $\Delta h := h_2 - h_1 \neq 0$. Let $k_1(x, y) = k_{\mathrm{RBF}}(x, y; h_1)$ and $k_2(x, y) = k_{\mathrm{RBF}}(x, y; h_2)$, and let $\pi(x) \propto \exp(-\alpha \exp(\frac{x^2}{2\beta}))$ be the target on $\mathcal{X} = \mathbb{R}$ where*

$$\alpha = \sqrt{\frac{h_2}{h_1}} \cdot \sqrt{\frac{\sigma^2}{\sigma^2 + \Delta h}} \cdot \frac{\sigma^2}{\Delta h}, \qquad \beta = \frac{\sigma^2(\sigma^2 + \Delta h)}{\Delta h}.$$

Then $\rho^*(x) = \mathcal{N}(x; 0, \sigma^2)$ is a fixed point of the h-SVGD mean field dynamics.

We emphasise that the $k_2 = ck_1$ case studied in the previous section is the focus of this paper due to its capacity to alleviate variance collapse. We leave a more detailed study of the general hybrid setting for future work.

4 Experiments

The problem of variance collapse in SVGD has been successfully mitigated in the setting of probabilistic graphical models where the conditional dependence structure enables π to be factorised and $\boldsymbol{R}(\,\cdot\,; k_2, \mu)$ to be replaced with a set of lower dimensional repulsive forces (Zhuo et al., 2018). Other methods such as S-SVGD (Gong et al., 2021) and GSVGD (Liu et al., 2022) have demonstrated that variance collapse can be avoided, at the expense of additional computational cost. In particular, S-SVGD requires computation of the optimal test directions, and GSVGD requires the projectors to be updated at each step. Our numerical experiments demonstrate that in the absence of a conditional dependence structure, and without incurring additional computational cost, h-SVGD can mitigate variance collapse while improving the inference capabilities of SVGD. In light of the discussion following Corollary 3.4, we argue that this mitigation occurs because the higher variance in the mean field limit offsets the variance underestimation in the finite particle setting. We measure variance collapse using dimension averaged marginal variance (DAMV), $\frac{1}{d}\sum_{j=1}^{d} \mathrm{Var}_j\left(\{\boldsymbol{x}_i\}_{i=1}^N\right)$, as is standard in the literature (Zhuo et al., 2018; Ba et al., 2019; 2021; Gong et al., 2021). Further details on the the experimental details can be found in Appendix B.

4.1 MULTIVARIATE GAUSSIAN MIXTURE

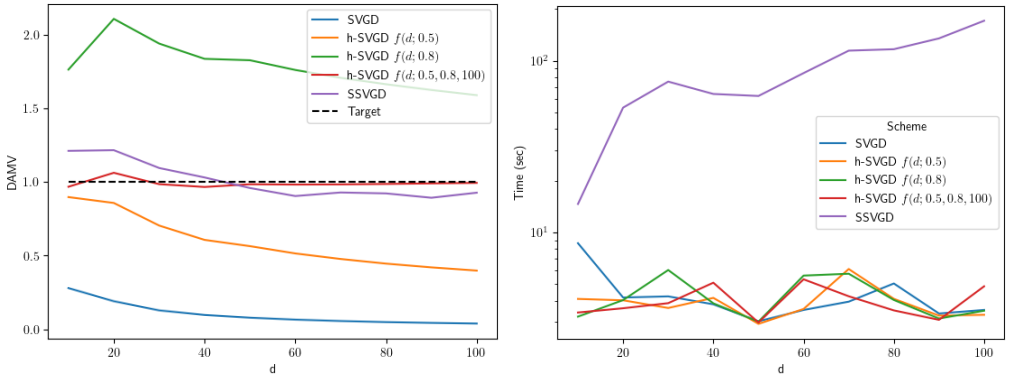
In this first example, we sample from a high-dimensional Gaussian distribution, a setting which has been explored previously to study variance collapse (Gong et al., 2021; Liu et al., 2022; Zhuo et al., 2018). We use a target distribution $\pi = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ for dimensions up to $d = 100$. We choose to sample $N = 50$ particles in order to demonstrate the performance of h-SVGD when d is much greater than N, as is often the case in high dimensional Bayesian inference. Particles are initialised from $\mathcal{N}(2\mathbf{1}_d, 2\boldsymbol{I}_d)$, where $\mathbf{1}_d$ is the vector of ones, and each SVGD variant is run for 2000 iterations with an initial step size of $\epsilon = 0.01$, adapted using AdaGrad. We run SVGD and S-SVGD with the RBF kernel k_{RBF} and compare against h-SVGD with kernels $k_1 = k_{\mathrm{RBF}}$ and $k_2 = f(d)k_{\mathrm{RBF}}$ for the following choices of factor f,

$$f(d;\alpha) := d^{\alpha}, \qquad \alpha \in (0,1), \tag{27}$$

$$f(d;\alpha_1,\alpha_2,c) := \frac{cd^{\alpha_1} + d^{1+\alpha_2}}{c+d}, \qquad \alpha_1,\alpha_2 \in (0,1), c > 0. \tag{28}$$

All algorithms use an adaptive bandwidth $h = \mathrm{med}^2/\log(N)$ where med is the median pairwise distance between particles (Liu & Wang, 2016). For S-SVGD, there is one bandwidth per dimension, so the median distances are computed along each projection, as described in (Gong et al., 2021). All configurations are run 5 times and results for each configuration are averaged.

Figure 1 demonstrates that h-SVGD provides a noticeable uplift in marginal variance estimation under the repulsive factors $f(d;0.5)$ and $f(d;0.8)$ when compared to SVGD, although increasing α too much leads to overestimation, especially in lower dimensions. The motivation for (28) was the observation that the variance estimation of (27) is better in lower dimensions for smaller α, and better in higher dimensions for larger α. Since $f(d;\alpha_1,\alpha_2,c) = \frac{c}{c+d}f(d;\alpha_1) + \frac{d}{c+d}f(d;\alpha_2)$, this choice should enable the advantages of both small and large α for moderately large c. Figure 1 demonstrates that choosing $\alpha_1 = 0.5$, $\alpha_2 = 0.8$ and $c = 100$ provides consistent variance estimation up to $d = 100$. An interesting avenue for future research would be to develop a more principled method of choosing such a a repulsive factor. The techniques in (Zhuo et al., 2018, Propositions 1-2) offer a promising direction. We remark that S-SVGD estimates the variance fairly consistently as the dimension increases. However, this comes at a significant increase in runtime, whereas there is no noticeable difference between the runtimes of SVGD and h-SVGD.



(a) DAMV (b) Time (seconds)

Figure 1: DAMV and runtime for different SVGD variants sampling from $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$.

4.2 BAYESIAN NEURAL NETWORK

In this section, we sample weights from a Bayesian neural network. Aside from scaling the repulsive kernel by $f(d;0.5,0.8,100)$, as defined in (28) during the previous experiment, our setup is identical to (Liu & Wang, 2016). For completeness, details are included in Appendix B. Table 1 shows that with no additional computational cost, the problem of variance collapse is mitigated under h-SVGD through an increased DAMV. Following Corollary 3.4, one may interpret the increased

variance in the mean field limit as an offset to the variance underestimation in finite dimensions, leading to improved variance estimation under h-SVGD. Table 2 demonstrates that it remains competitive at inference through improved test log-likelihood (LL) for all datasets and improved root mean squared error (RMSE) for all but one dataset. We observe that for two datasets, Kin8nm and Protein, h-SVGD enjoys an improvement in RMSE and LL despite a small decrease in the DAMV. This suggests that the improved performance may be linked to another property of h-SVGD. We leave further exploration of such properties for future work. Appendix B includes a comparison of the same metrics between h-SVGD and S-SVGD to emphasise that the improvement in variance estimation comes at a lower cost.

Table 1: DAMV and runtime in seconds of SVGD and h-SVGD.

	DAMV		Runtime (seconds)	
Dataset	**SVGD**	**h-SVGD**	**SVGD**	**h-SVGD**
Boston	0.051 ± 0.011	$\mathbf{0.087 \pm 0.010}$	28.9 ± 1.4	$\mathbf{28.5 \pm 0.9}$
Concrete	0.084 ± 0.01	$\mathbf{0.102 \pm 0.006}$	28.7 ± 1.3	28.7 ± 1.3
Energy	0.065 ± 0.015	$\mathbf{0.106 \pm 0.011}$	30.8 ± 2.8	$\mathbf{30.7 \pm 2.2}$
Kin8nm	$\mathbf{0.105 \pm 0.003}$	0.093 ± 0.003	$\mathbf{35.9 \pm 1.3}$	36.0 ± 1.3
Naval	0.059 ± 0.004	$\mathbf{0.068 \pm 0.011}$	$\mathbf{33.4 \pm 0.8}$	34.5 ± 0.9
Combined	0.128 ± 0.008	$\mathbf{0.138 \pm 0.006}$	$\mathbf{36.3 \pm 1.5}$	36.7 ± 2.4
Protein	$\mathbf{0.089 \pm 0.001}$	0.084 ± 0.001	$\mathbf{72.7 \pm 1.0}$	72.8 ± 1.7
Wine	0.068 ± 0.005	$\mathbf{0.075 \pm 0.005}$	$\mathbf{29.5 \pm 1.4}$	29.8 ± 1.1
Yacht	0.060 ± 0.020	$\mathbf{0.121 \pm 0.012}$	$\mathbf{29.3 \pm 0.4}$	29.8 ± 1.3
Year	$0.011 \pm \mathrm{NA}$	$\mathbf{0.012 \pm NA}$	$673 \pm \mathrm{NA}$	$\mathbf{666 \pm NA}$

Table 2: Average RMSE and LL of SGVD and h-SVGD evaluated on the test dataset.

	Test RMSE		Test LL	
Dataset	**SVGD**	**h-SVGD**	**SVGD**	**h-SVGD**
Boston	3.094 ± 0.579	$\mathbf{3.001 \pm 0.584}$	-2.123 ± 0.116	$\mathbf{-1.988 \pm 0.221}$
Concrete	5.857 ± 0.468	$\mathbf{5.210 \pm 0.529}$	-2.616 ± 0.099	$\mathbf{-2.535 \pm 0.179}$
Energy	1.528 ± 0.169	$\mathbf{1.040 \pm 0.128}$	-1.702 ± 0.094	$\mathbf{-0.805 \pm 0.104}$
Kin8nm	0.124 ± 0.005	$\mathbf{0.090 \pm 0.003}$	-1.293 ± 0.108	$\mathbf{0.468 \pm 0.090}$
Naval	0.006 ± 0.000	$\mathbf{0.004 \pm 0.000}$	-1.353 ± 0.161	$\mathbf{-0.090 \pm 0.105}$
Combined	4.105 ± 0.220	$\mathbf{4.057 \pm 0.218}$	-2.459 ± 0.051	$\mathbf{-2.354 \pm 0.052}$
Protein	4.791 ± 0.025	$\mathbf{4.600 \pm 0.026}$	-2.633 ± 0.035	$\mathbf{-2.456 \pm 0.017}$
Wine	0.637 ± 0.044	$\mathbf{0.626 \pm 0.045}$	-1.463 ± 0.120	$\mathbf{-0.750 \pm 0.097}$
Yacht	$\mathbf{1.677 \pm 0.571}$	1.861 ± 0.662	-1.587 ± 0.120	$\mathbf{-0.813 \pm 0.227}$
Year	$8.837 \pm \mathrm{NA}$	$\mathbf{8.689 \pm NA}$	$-2.883 \pm \mathrm{NA}$	$\mathbf{-2.872 \pm NA}$

5 Conclusion

We developed the mean field theory of h-SVGD by proving the existence of a solution to the hybrid Stein PDE and identifying it as a gradient flow on a functional other than the KL divergence. We characterised the mean field fixed point for the special case $k_2 = ck_1$ and demonstrated that h-SVGD does not converge to the target in the mean field limit. We provided a result on the dissipation of the new functional, as well as a discrete time version, otherwise known as a descent lemma. We also highlighted the complexities of the gradient flow in the general hybrid kernel setting. Experimental results demonstrated that h-SVGD can alleviate variance collapse in high dimensions at a much lower cost than other SVGD variants. We also showed that h-SVGD maintains its performance on high dimensional inference tasks, whilst improving variance estimation without the additional computational cost required of other SVGD variants. One interesting direction for future research is to find a principled method of scaling the repulsive kernel, using the theoretical analysis in (Zhuo et al., 2018) as a starting point. Another avenue is to further develop the theory of h-SVGD in the general hybrid kernel setting.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Jimmy Ba, Murat A Erdogdu, Marzyeh Ghassemi, Taiji Suzuki, Shengyang Sun, Denny Wu, and Tianzong Zhang. Towards characterizing the high-dimensional bias of kernel-based particle inference algorithms. In *Second Symposium on Advances in Approximate Bayesian Inference*, 2019.

Jimmy Ba, Murat A Erdogdu, Marzyeh Ghassemi, Shengyang Sun, Taiji Suzuki, Denny Wu, and Tianzong Zhang. Understanding the variance collapse of SVGD in high dimensions. In *International Conference on Learning Representations*, 2021.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Peng Chen and Omar Ghattas. Projected Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 33:1947–1958, 2020.

Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, and Philippe Rigollet. SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. *Advances in Neural Information Processing Systems*, 33:2098–2109, 2020.

Francesco D'Angelo and Vincent Fortuin. Annealed Stein variational gradient descent. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021.

Francesco D'Angelo, Vincent Fortuin, and Florian Wenzel. On Stein variational neural network ensembles. *arXiv preprint arXiv:2106.10760*, 2021.

Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of Stein variational gradient descent. *Journal of Machine Learning Research*, 24:1–39, 2023.

Wenbo Gong, Yingzhen Li, and José Miguel Hernández-Lobato. Sliced kernelized Stein discrepancy. In *International Conference on Learning Representations*, 2021.

Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton. A non-asymptotic analysis for Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 33:4672–4682, 2020.

Chang Liu and Jun Zhu. Riemannian Stein variational gradient descent for Bayesian inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Qiang Liu. Stein variational gradient descent as gradient flow. *Advances in Neural Information Processing Systems*, 30, 2017.

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

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

Xing Liu, Harrison Zhu, Jean-François Ton, George Wynne, and Andrew Duncan. Grassmann Stein variational gradient descent. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 2002–2021. PMLR, 2022.

Jianfeng Lu, Yulong Lu, and James Nolen. Scaling limit of the Stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671, 2019.

Adil Salim, Lukang Sun, and Peter Richtarik. A convergence theory for SVGD in the population limit under Talagrand's inequality T1. In *International Conference on Machine Learning*, pp. 19139–19152. PMLR, 2022.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Dilin Wang, Zhe Zeng, and Qiang Liu. Stein variational message passing for continuous graphical models. In *International Conference on Machine Learning*, pp. 5219–5227. PMLR, 2018.

Dilin Wang, Ziyang Tang, Chandrajit Bajaj, and Qiang Liu. Stein variational gradient descent with matrix-valued kernels. *Advances in Neural Information Processing Systems*, 32, 2019.

Jingwei Zhuo, Chang Liu, Jiaxin Shi, Jun Zhu, Ning Chen, and Bo Zhang. Message passing Stein variational gradient descent. In *International Conference on Machine Learning*, pp. 6018–6027. PMLR, 2018.

A PROOFS

Lemma A.1. *Under assumptions (**A**1), (**A**4), (**B**1), (**B**2) the map*

$$(\boldsymbol{z}, \mu) \mapsto E(\boldsymbol{z}, \mu) := \int_{\mathcal{X}} -k_1(\boldsymbol{x}, \boldsymbol{z}) \nabla V(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} k_2(\boldsymbol{x}, \boldsymbol{z}) d\mu(\boldsymbol{x})$$

is L-Lipschitz. That is,

$$\|E(\boldsymbol{z}, \mu) - E(\boldsymbol{z}', \mu')\|_2 \leq L(\|\boldsymbol{z} - \boldsymbol{z}'\|_2 + W_2(\mu, \mu'))$$

where $L > 0$ depends on k_1, k_2 and V.

Proof. Largely following the proof of Lemma 14 (Korba et al., 2020), choosing an optimal coupling γ of μ and μ',

$$\begin{aligned}
\|E(\boldsymbol{z}, \mu) - E(\boldsymbol{z}', \mu')\|_2 \leq \Big\| & \mathbb{E}_\gamma \left[\nabla V(\boldsymbol{x})(k_1(\boldsymbol{x}, \boldsymbol{z}) - k_1(\boldsymbol{x}', \boldsymbol{z}'))\right] \\
& + \mathbb{E}_\gamma \left[(\nabla V(\boldsymbol{x}') - \nabla V(\boldsymbol{x})) k_1(\boldsymbol{x}', \boldsymbol{z}')\right] \\
& + \mathbb{E}_\gamma \left[\nabla_{\boldsymbol{x}} k_2(\boldsymbol{x}, \boldsymbol{z}) - \nabla_{\boldsymbol{x}'} k_2(\boldsymbol{x}', \boldsymbol{z}')\right] \Big\| \\
& \leq D\mathbb{E}_\gamma \left[\|\boldsymbol{x} - \boldsymbol{x}'\|_2 + \|\boldsymbol{z} - \boldsymbol{z}'\|_2\right] \\
& + BM\mathbb{E}_\gamma \left[\|\boldsymbol{x} - \boldsymbol{x}'\|_2\right] \\
& + D\mathbb{E}_\gamma \left[\|\boldsymbol{x} - \boldsymbol{x}'\|_2 + \|\boldsymbol{z} - \boldsymbol{z}'\|_2\right] \\
& \leq (2D + BM) \left(\|\boldsymbol{z} - \boldsymbol{z}'\|_2 + W_2(\mu, \mu')\right).
\end{aligned}$$

Note that the second term is bounded using the relaxed Assumption (**B**2) and there is no need to require that $|V|$ is bounded by a constant. □

Proof of Proposition 3.1. This follows identically to the proof of (Korba et al., 2020, Proposition 7) with Lemma A.1 in place of Lemma 14. □

Proof of Proposition 3.2. The proof largely follows those of (Lu et al., 2019, Theorems 2.4 and 3.2) with some minor adjustments. Notably, after fixing $r > 0$ and defining

$$Y_r := \left\{ u \in Y : \sup_{\boldsymbol{x} \in \mathcal{X}} |u(\boldsymbol{x}) - \boldsymbol{x}| < r \right\}$$

and the complete metric space

$$\begin{aligned}
S_r &:= C([0, T_0]; Y_r), \\
d_S(u, v) &:= \sup_{t \in [0, T_0]} d_Y(u(t), v(t))
\end{aligned}$$

for some sufficiently small T_0 (to be determined later), the operator $\mathcal{G} : u(t, \cdot) \mapsto \mathcal{G}(u)(t, \cdot)$ must be modified to act on $u \in S_r$ via

$$\begin{aligned}
\mathcal{G}(u)(t, \boldsymbol{x}) = \boldsymbol{x} & - \int_0^t \int_{\mathcal{X}} \nabla K_2(u(s, \boldsymbol{x}) - u(s, \boldsymbol{x}')) \nu(d\boldsymbol{x}') ds \\
& - \int_0^t \int_{\mathcal{X}} K_1(u(s, \boldsymbol{x}) - u(s, \boldsymbol{x}')) \nabla V(u(s, \boldsymbol{x}')) \nu(d\boldsymbol{x}') ds.
\end{aligned}$$

Note that we use $\mathcal{G}$ instead of the $\mathcal{F}$ used in (Lu et al., 2019) to avoid confusion between the functional $\mathcal{F}$ defined in (17). The same techniques of (Lu et al., 2019) are sufficient to establish the required bounds to show that $\mathcal{G}$ is a contraction on S_r for sufficiently small T_0. Note that Assumptions (**B**1), (**A**2) and (**A**3) are used to establish this. So the unique fixed point $X(\cdot, \cdot; \nu) \in S_r$ of $\mathcal{G}$ solves (13) in the interval $[0, T_0]$.

The $\min\left(\|\nabla K_1\|_\infty, \|\nabla K_2\|_\infty\right)$ term emerges because the telescoping in (Lu et al., 2019, Equation (3.8)) can be performed with either kernel. The remainder of the proof follows (Lu et al., 2019, Theorems 2.4 and 3.2). □

Proof of Proposition 3.3. Following (Santambrogio, 2015, Definition 7.12), a functional derivative of $\mathcal{F}$ is a measurable function $\frac{\delta \mathcal{F}}{\delta \rho}(\rho)$ satisfying

$$\frac{d}{d\epsilon}\mathcal{F}(\rho + \epsilon\chi)\bigg|_{\epsilon=0} = \int \frac{\delta \mathcal{F}}{\delta \rho}(\rho) d\chi$$

for all perturbations $\chi = \tilde{\rho} - \rho$ with $\tilde{\rho} \in L_c^\infty(\mathcal{X}) \cap \mathcal{P}(\mathcal{X})$. If it exists, $\frac{\delta \mathcal{F}}{\delta \rho}(\rho)$ is unique up to additive constants.

We first compute

$$\begin{aligned}
&\frac{d}{d\epsilon}\mathcal{F}(\rho + \epsilon\chi)\bigg|_{\epsilon=0} \\
&= \frac{d}{d\epsilon}\bigg(\int c \log(\rho(\boldsymbol{x}) + \epsilon\chi(\boldsymbol{x}))\rho(\boldsymbol{x}) d\boldsymbol{x} + \epsilon \int c \log(\rho(\boldsymbol{x}) + \epsilon\chi(\boldsymbol{x}))\chi(\boldsymbol{x}) d\boldsymbol{x} \\
&\quad - \int \log \pi(\boldsymbol{x})\rho(\boldsymbol{x}) d\boldsymbol{x} - \epsilon \int \log \pi(\boldsymbol{x})\chi(\boldsymbol{x}) d\boldsymbol{x} \bigg)\bigg|_{\epsilon=0} \\
&= \bigg(\int c \frac{\chi(\boldsymbol{x})}{\rho(\boldsymbol{x}) + \epsilon\chi(\boldsymbol{x})}\rho(\boldsymbol{x}) d\boldsymbol{x} + \int c \log(\rho(\boldsymbol{x}) + \epsilon\chi(\boldsymbol{x}))\chi(\boldsymbol{x}) d\boldsymbol{x} \\
&\quad + \epsilon \int c \frac{\chi(\boldsymbol{x})}{\rho(\boldsymbol{x}) + \epsilon\chi(\boldsymbol{x})}\chi(\boldsymbol{x}) d\boldsymbol{x} - \int \log \pi(\boldsymbol{x})\chi(\boldsymbol{x}) d\boldsymbol{x} \bigg)\bigg|_{\epsilon=0} \\
&= \int c \log \rho(\boldsymbol{x}) - \log \pi(\boldsymbol{x})\chi(\boldsymbol{x}) d\boldsymbol{x} + c \int \chi(\boldsymbol{x}) d\boldsymbol{x}. \qquad (29)
\end{aligned}$$

Since $\rho, \tilde{\rho} \in \mathcal{P}(\mathcal{X})$, the final integral is zero. So the functional gradient of $\mathcal{F}$ is

$$\frac{\delta \mathcal{F}}{\delta \rho}(\rho) = c \log \rho - \log \pi.$$

Its Wasserstein gradient is then

$$\nabla_W \mathcal{F}(\rho) = c\nabla \log \rho - \nabla \log \pi.$$

Since $k_2 = ck_1$, equation (10) can be written as

$$\begin{aligned}
\frac{d\boldsymbol{x}}{dt} &= \int \big(k_1(\boldsymbol{x}', \boldsymbol{x})\nabla_{\boldsymbol{x}'} \log \pi(\boldsymbol{x}') + c\nabla_{\boldsymbol{x}'} k_1(\boldsymbol{x}', \boldsymbol{x})\big)\rho(\boldsymbol{x}') d\boldsymbol{x}' \\
&= \int k_1(\boldsymbol{x}', \boldsymbol{x})\nabla_{\boldsymbol{x}'} \log \pi(\boldsymbol{x}')\rho(\boldsymbol{x}') - ck_1(\boldsymbol{x}', \boldsymbol{x})\nabla_{\boldsymbol{x}'}\rho(\boldsymbol{x}') d\boldsymbol{x}' \\
&= \int k_1(\boldsymbol{x}', \boldsymbol{x})\nabla_{\boldsymbol{x}'} \log \pi(\boldsymbol{x}')\rho(\boldsymbol{x}') - ck_1(\boldsymbol{x}', \boldsymbol{x})\nabla_{\boldsymbol{x}'} \log \rho(\boldsymbol{x}')\rho(\boldsymbol{x}') d\boldsymbol{x}' \\
&= T_{k_1,\rho}(\nabla \log \pi - c\nabla \log \rho)(\boldsymbol{x}) \\
&= -T_{k_1,\rho}(\nabla_W \mathcal{F}(\rho))(\boldsymbol{x}).
\end{aligned}$$

The continuity equation $\partial_t \rho_t + \nabla \cdot \left(\rho_t \frac{d\boldsymbol{x}}{dt}\right) = 0$ then becomes

$$\begin{aligned}
\partial_t \rho_t &= \nabla \cdot (\rho_t T_{k_1,\rho}(\nabla_W \mathcal{F}(\rho))(\boldsymbol{x})) \\
&= \nabla \cdot \left(\rho_t T_{k_1,\rho_t}\left(c\nabla \log \rho_t - \nabla \log \pi\right)\right).
\end{aligned}$$

□

Proof of Corollary (3.4). A measure ρ^* satisfying $c\nabla \log \rho^* = \nabla \log \pi$ will be a fixed point because this would imply $\nabla_W \mathcal{F}(\rho) = 0$. Solving this for ρ^*, we have

$$\begin{aligned}
c\nabla \log \rho^*(\boldsymbol{x}) &= \nabla \log \pi(\boldsymbol{x}) \\
c \log \rho^*(\boldsymbol{x}) &= \log \pi(\boldsymbol{x}) + A \\
\log(\rho^*(\boldsymbol{x})^c) &= \log(e^A \pi(\boldsymbol{x})) \\
\rho^*(\boldsymbol{x})^c &= e^A \pi(\boldsymbol{x}) \\
\rho^*(\boldsymbol{x}) &= e^{A/c} \pi(\boldsymbol{x})^{1/c}
\end{aligned}$$

for some $A \in \mathbb{R}$.

□

Proof of Proposition 3.5. Using (19), integration by parts, and the fact that $T_{k_1,\rho} : L^2(\rho_t)^d \to \mathcal{H}_1^d$ is the adjoint of the inclusion $\imath : \mathcal{H}_1^d \to L^2(\rho_t)^d$,

$$\begin{aligned}
\frac{d}{dt}\mathcal{F}(\rho_t) &= \frac{d}{dt}\int \rho_t(\boldsymbol{x})\left(c\log\rho_t(\boldsymbol{x}) - \log\pi(\boldsymbol{x})\right)d\boldsymbol{x} \\
&= \int \left(c\log\rho_t(\boldsymbol{x}) - \log\pi(\boldsymbol{x})\right)\frac{\partial}{\partial t}\rho_t(\boldsymbol{x}) + c\rho_t(\boldsymbol{x})\frac{\partial}{\partial t}\log\rho_t(\boldsymbol{x})d\boldsymbol{x} \\
&= \int \left(c\log\rho_t(\boldsymbol{x}) - \log\pi(\boldsymbol{x})\right)\nabla\cdot\left(\rho_t(\boldsymbol{x})T_{k_1,\rho_t}(c\nabla\log\rho_t(\boldsymbol{x}) - \nabla\log\pi)(\boldsymbol{x})\right)d\boldsymbol{x} \\
&\quad + \int c\rho_t(\boldsymbol{x})\frac{\partial}{\partial t}\log\rho_t(\boldsymbol{x})d\boldsymbol{x} \\
&= -\int \nabla\left(c\log\rho_t(\boldsymbol{x}) - \log\pi(\boldsymbol{x})\right)\cdot T_{k_1,\rho_t}(c\nabla\log\rho_t - \nabla\log\pi)(\boldsymbol{x})\rho_t(\boldsymbol{x})d\boldsymbol{x} \\
&\quad + \int c\rho_t(\boldsymbol{x})\frac{\partial}{\partial t}\log\rho_t(\boldsymbol{x})d\boldsymbol{x} \\
&= -\left\langle c\nabla\log\rho_t - \nabla\log\pi, T_{k_1,\rho_t}(c\nabla\log\rho_t - \nabla\log\pi)\right\rangle_{L^2(\rho_t)^d} \\
&\quad + \int c\rho_t(\boldsymbol{x})\frac{\partial}{\partial t}\log\rho_t(\boldsymbol{x})d\boldsymbol{x} \\
&= -\left\|T_{k_1,\rho_t}(c\nabla\log\rho_t - \nabla\log\pi)\right\|_{\mathcal{H}_1^d}^2 + c\int \rho_t(\boldsymbol{x})\frac{\partial}{\partial t}\log\rho_t(\boldsymbol{x})d\boldsymbol{x} \\
&= -c^2\left\|T_{k_1,\rho_t}(\nabla\log\rho_t - \nabla\log\rho^*)\right\|_{\mathcal{H}_1^d}^2 + c\int \rho_t(\boldsymbol{x})\frac{\partial}{\partial t}\log\rho_t(\boldsymbol{x})d\boldsymbol{x}. \qquad (30)
\end{aligned}$$

Also,

$$\begin{aligned}
\mathcal{F}(\rho_t) &= \mathbb{E}_{\boldsymbol{x}\sim\rho_t}\left[c\log\rho_t(\boldsymbol{x}) - \log\pi(\boldsymbol{x})\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\rho_t}\left[c\log\rho_t(\boldsymbol{x}) - \log\left(A\rho^*(\boldsymbol{x})^c\right)\right] \\
&= c\mathbb{E}_{\boldsymbol{x}\sim\rho_t}\left[\log\rho_t(\boldsymbol{x}) - \log\rho^*(\boldsymbol{x})\right] + \log(A) \\
&= c\mathrm{KL}(\rho_t \parallel \rho^*) + \log(A) \qquad (31)
\end{aligned}$$

for some $A \in \mathbb{R}$, we have $\frac{1}{c}\frac{d}{dt}\mathcal{F}(\rho_t) = \frac{d}{dt}\mathrm{KL}(\rho_t \parallel \rho^*)$. Therefore,

$$\begin{aligned}
\frac{d}{dt}\mathrm{KL}(\rho_t \parallel \rho^*) &= -\frac{1}{c}\left\|T_{k_1,\rho_t}(c\nabla\log\rho_t - \nabla\log\pi)\right\|_{\mathcal{H}_1^d}^2 + \int \rho_t(\boldsymbol{x})\frac{\partial}{\partial t}\log\rho_t(\boldsymbol{x})d\boldsymbol{x} \\
&= -c\left\|T_{k_1,\rho_t}(\nabla\log\rho_t - \nabla\log\pi^{1/c})\right\|_{\mathcal{H}_1^d}^2 + \int \rho_t(\boldsymbol{x})\frac{\partial}{\partial t}\log\rho_t(\boldsymbol{x})d\boldsymbol{x} \\
&= -c\left\|T_{k_1,\rho_t}(\nabla\log\rho_t - \nabla\log\rho^*)\right\|_{\mathcal{H}_1^d}^2 + \int \rho_t(\boldsymbol{x})\frac{\partial}{\partial t}\log\rho_t(\boldsymbol{x})d\boldsymbol{x}.
\end{aligned}$$

To simplify notation in the calculations below, set $u = T_{k_1,\rho_t}(c\nabla\log\rho_t)$ and $v = T_{k_1,\rho_t}(\nabla\log\pi)$. Recall the identity

$$\langle u, u - v\rangle = \frac{1}{2}\left(\|u\|^2 + \|u - v\|^2 - \|v\|^2\right). \qquad (32)$$

Now we apply the multivariate chain rule to the remainder term along with the fact about the adjoint of T_{k_1,ρ_t}, the identity in (32), and the triangle inequality. This gives

$$\begin{aligned}
\int \rho_t(\boldsymbol{x}) \frac{\partial}{\partial t} \log \rho_t(\boldsymbol{x}) d\boldsymbol{x} &= \int \rho_t(\boldsymbol{x}) \nabla \log \rho_t(\boldsymbol{x}) \cdot \frac{d\boldsymbol{x}}{dt} d\boldsymbol{x} \\
&= -\int \rho_t(\boldsymbol{x}) \nabla \log \rho_t(\boldsymbol{x}) \cdot T_{k_1,\rho_t}(c\nabla \log \rho_t - \nabla \log \pi)(\boldsymbol{x}) d\boldsymbol{x} \qquad (33) \\
&= -\left\langle \nabla \log \rho_t, T_{k_1,\rho_t}(c\nabla \log \rho_t - \nabla \log \pi) \right\rangle_{L^2(\rho_t)^d} \\
&= -\left\langle T_{k_1,\rho_t}(\nabla \log \rho_t), T_{k_1,\rho_t}(c\nabla \log \rho_t - \nabla \log \pi) \right\rangle_{\mathcal{H}_1^d} \\
&= -\frac{1}{c} \left\langle u, u - v \right\rangle_{\mathcal{H}_1^d} \\
&= \frac{1}{2c} \left(-\|u\|_{\mathcal{H}_1^d}^2 + \|v\|_{\mathcal{H}_1^d}^2 - \|u - v\|_{\mathcal{H}_1^d}^2 \right) \\
&\leq \frac{1}{2c} \left(-\|u\|_{\mathcal{H}_1^d}^2 + \|u - v\|_{\mathcal{H}_1^d}^2 + \|u\|_{\mathcal{H}_1^d}^2 - \|u - v\|_{\mathcal{H}_1^d}^2 \right) \\
&= 0.
\end{aligned}$$

The final statement that $\frac{d}{dt}\mathcal{F}(\rho^*) = 0$ follows from equations (30) and (33) by observing that $c\nabla \log \rho^* - \nabla \log \pi = 0$. □

Proof of Proposition 3.6. This follows from applying Theorem 3.3 (Liu, 2017) with ρ^* instead of the target, then substituting in (31). □

Proof of Proposition 3.7. To recover the continuity equation, apply the convolution theorem to (22)

$$\begin{aligned}
\mathscr{F}^{-1}\left(\frac{\mathscr{F}(K_2)}{\mathscr{F}(K_1)} \cdot \mathscr{F}(\nabla \rho) \right)(x) &= r(x;\rho) \\
\mathscr{F}(K_2)(\omega) \cdot \mathscr{F}(\nabla \rho)(\omega) &= \mathscr{F}(K_1)(\omega) \cdot \mathscr{F}(r)(\omega) \\
\mathscr{F}(K_2 * \nabla \rho)(\omega) &= \mathscr{F}(K_1 * r)(\omega) \\
(K_2 * \nabla \rho)(x) &= (K_1 * r)(x) \\
\int K_2(x - y) \nabla \log \rho(y) \rho(y) dy &= \int K_1(x - y) \frac{r(y)}{\rho(y)} \rho(y) dy \\
\left(T_{k_2,\rho} \nabla \log \rho\right)(x) &= \left(T_{k_1,\rho}(r/\rho)\right)(x)
\end{aligned}$$

Equation (10) can now be rewritten as

$$\begin{aligned}
\frac{dx}{dt} &= \int \left(k_1(x', x) \nabla_{x'} \log \pi(x') + \nabla_{x'} k_2(x', x) \right) \rho(x') dx' \\
&= \int k_1(x', x) \nabla_{x'} \log \pi(x') \rho(x') - k_2(x', x) \nabla_{x'} \rho(x') dx' \\
&= \int \left(k_1(x', x) \nabla_{x'} \log \pi(x') - k_2(x', x) \nabla_{x'} \log \rho(x') \right) \rho(x') dx' \\
&= \left(T_{k_1,\rho} \nabla \log \pi\right)(x) - \left(T_{k_2,\rho} \nabla \log \rho\right)(x) \\
&= \left(T_{k_1,\rho} \left(\nabla \log \pi - \nabla R(\cdot\,; \rho)\right)\right)(x),
\end{aligned}$$

where $\nabla R(x;\rho) := \frac{r(x;\rho)}{\rho(x)}$, yielding the continuity equation (23).

As in the proof of Proposition 3.3, let $\chi = \tilde{\rho} - \rho$ with $\tilde{\rho} \in L_c^\infty(\mathcal{X}) \cap \mathcal{P}(\mathcal{X})$. We first compute

$$
\begin{aligned}
&\frac{d}{d\epsilon}\mathcal{F}(\rho + \epsilon\chi)\bigg|_{\epsilon=0} \\
&= \frac{d}{d\epsilon}\bigg(\int R(x; \rho + \epsilon\chi) d\rho(x) + \epsilon \int R(x; \rho + \epsilon\chi) d\chi(x) \\
&\quad - \int \log \pi(x) d\rho(x) - \epsilon \int \log \pi(x) d\chi(x) \bigg)\bigg|_{\epsilon=0} \\
&= \bigg(\frac{d}{d\epsilon} \int R(x; \rho + \epsilon\chi) d\rho(x) + \int R(x; \rho + \epsilon\chi) d\chi(x) + \epsilon \frac{d}{d\epsilon} \int R(x; \rho + \epsilon\chi) d\chi(x) \\
&\quad - \int \log \pi(x) d\chi(x) \bigg)\bigg|_{\epsilon=0} \\
&= \int R(x; \rho) - \log \pi(x) d\chi(x) \\
&\quad + \bigg(\int \frac{\partial}{\partial \epsilon} R(x; \rho + \epsilon\chi) d\rho(x) + \epsilon \int \frac{\partial}{\partial \epsilon} R(x; \rho + \epsilon\chi) d\chi(x) \bigg) \bigg|_{\epsilon=0} \\
&= \int R(x; \rho) - \log \pi(x) d\chi(x) + \int \frac{\partial}{\partial \epsilon} R(x; \rho + \epsilon\chi) d(\rho + \epsilon\chi)(x) \bigg|_{\epsilon=0}. \qquad (34)
\end{aligned}
$$

By assumption, the remainder term above is equal to zero and the functional derivative of $\mathcal{F}$ is therefore

$$
\frac{\delta \mathcal{F}}{\delta \rho}(\rho) = R_1(x; \rho) - \log \pi(x),
$$

so its Wasserstein gradient is

$$
\begin{aligned}
\nabla_W \mathcal{F}(\rho) &= \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\rho) \\
&= R(x; \rho) - \nabla \log \pi(x).
\end{aligned}
$$

□

Proof of Proposition 3.8. A direct computation of (22) along with the definitions of r and R yields

$$
\begin{aligned}
\frac{\mathscr{F}(K_2)(\omega)}{\mathscr{F}(K_1)(\omega)} \cdot \mathscr{F}(\nabla \rho^*)(\omega) &= \sqrt{\frac{h_2}{h_1}} \cdot \exp(-2\pi^2\omega^2\Delta h) \cdot 2\pi i\omega \exp(-2\pi^2\omega^2\sigma^2) \\
&= \sqrt{\frac{h_2}{h_1}} \cdot 2\pi i\omega \cdot \exp(-2\pi^2\omega^2(\sigma^2 + \Delta h)) \\
r(x; \rho^*) &= -\sqrt{\frac{h_2}{h_1}} \cdot \frac{1}{\sqrt{2\pi}} \cdot \frac{x}{(\sigma^2 + \Delta h)^{3/2}} \cdot \exp\left(-\frac{x^2}{2(\sigma^2 + \Delta h)} \right) \\
\nabla R(x; \rho^*) &= -\sqrt{\frac{h_2}{h_1}} \cdot \frac{\sigma}{(\sigma^2 + \Delta h)^{3/2}} \cdot x \exp\left(-\frac{x^2}{2(\sigma^2 + \Delta h)} \right) \cdot \exp\left(\frac{x^2}{2\sigma^2} \right) \\
&= -\sqrt{\frac{h_2}{h_1}} \cdot \frac{\sigma}{(\sigma^2 + \Delta h)^{3/2}} \cdot x \exp\left(-\frac{\Delta h x^2}{2\sigma^2(\sigma^2 + \Delta h)} \right).
\end{aligned}
$$

Now computing $\nabla \log \pi$, using some $A \in \mathbb{R}$ for the normalising constant, we have

$$\begin{aligned}
\pi(x) &= A \exp\left(-\alpha \exp\left(\frac{x^2}{2\beta}\right)\right) \\
\log \pi(x) &= \log(A) - \alpha \exp\left(\frac{x^2}{2\beta}\right) \\
\nabla \log \pi(x) &= -\frac{\alpha x}{\beta} \exp\left(\frac{x^2}{2\beta}\right) \\
&= -\sqrt{\frac{h_2}{h_1}} \cdot \frac{\sigma}{(\sigma^2 + \Delta h)^{3/2}} \cdot x \exp\left(-\frac{\Delta h x^2}{2\sigma^2(\sigma^2 + \Delta h)}\right).
\end{aligned}$$

Since $\nabla R(x; \rho^*) = \nabla \log \pi(x)$, equation (23) implies that ρ^* is a fixed point. □

B Additional Experimental Results and Details

B.1 Bayesian Neural Network

The results presented in Section 4.2 follow the settings of (Liu & Wang, 2016). In particular, we use normal priors for the network weights and Gamma priors for the inverse covariances. There is one hidden layer with 50 units for most datasets, Protein and Year being the exceptions with 100 units each. The datasets are randomly partitioned into 90% for training and 10% for testing with results averaged over 20 trials, Protein and Year being the exceptions with 5 trials and 1 trial respectively. The number of particles in each case is 20, the activation function is $\mathrm{RELU}(x) = \max(0, x)$, the number of iterations is 2000, and the mini-batch size is 100 for all datasets except for Year, which uses a mini-batch size of 1000.

We recreate Tables 1 and 2 below, this time comparing h-SVGD against S-SVGD (Gong et al., 2021). Table 3 shows that h-SVGD outperforms S-SVGD in mitigation of variance collapse on all but one dataset. We note that h-SVGD achieves this with a significantly faster runtime for all datasets, which is due to S-SVGD requiring additional optimistaion of the projection matrix. Table 4 shows that S-SVGD and h-SVGD are comparable in both RMSE and LL metrics. The h-SVGD algorithm achieves a better LL score on more datasets, but S-SVGD achieves a better RMSE score on more datasets.

Table 3: DAMV and runtime in seconds of S-SVGD and h-SVGD.

	DAMV		Runtime (seconds)	
Dataset	**S-SVGD**	**h-SVGD**	**S-SVGD**	**h-SVGD**
Boston	0.035 ± 0.002	$\mathbf{0.087 \pm 0.010}$	208 ± 13	$\mathbf{28.5 \pm 0.9}$
Concrete	0.070 ± 0.004	$\mathbf{0.102 \pm 0.006}$	148 ± 56	$\mathbf{28.7 \pm 1.3}$
Energy	0.053 ± 0.005	$\mathbf{0.106 \pm 0.011}$	156 ± 28	$\mathbf{30.7 \pm 2.2}$
Kin8nm	0.083 ± 0.002	$\mathbf{0.093 \pm 0.003}$	141 ± 3.9	$\mathbf{36.0 \pm 1.3}$
Naval	$\mathbf{0.070 \pm 0.021}$	0.068 ± 0.011	237 ± 11	$\mathbf{34.5 \pm 0.9}$
Combined	0.118 ± 0.005	$\mathbf{0.138 \pm 0.006}$	116 ± 18	$\mathbf{36.7 \pm 2.4}$
Protein	0.057 ± 0.006	$\mathbf{0.084 \pm 0.001}$	390 ± 20	$\mathbf{72.8 \pm 1.7}$
Wine	0.029 ± 0.002	$\mathbf{0.075 \pm 0.005}$	210 ± 10	$\mathbf{29.8 \pm 1.1}$
Yacht	0.066 ± 0.009	$\mathbf{0.121 \pm 0.012}$	97.9 ± 1.2	$\mathbf{30.0 \pm 1.3}$
Year	$0.012 \pm \mathrm{NA}$	$0.012 \pm \mathrm{NA}$	$12488 \pm \mathrm{NA}$	$\mathbf{666 \pm NA}$

Table 4: Average RMSE and LL of SSGVD and h-SVGD evaluated on the test dataset.

	Test RMSE		Test LL	
Dataset	**S-SVGD**	**h-SVGD**	**S-SVGD**	**h-SVGD**
Boston	3.024 ± 0.604	$\mathbf{3.001 \pm 0.584}$	-2.088 ± 0.322	$\mathbf{-1.988 \pm 0.221}$
Concrete	$\mathbf{5.073 \pm 0.522}$	5.210 ± 0.529	-2.563 ± 0.239	$\mathbf{-2.535 \pm 0.179}$
Energy	$\mathbf{0.923 \pm 0.123}$	1.040 ± 0.128	$\mathbf{-0.631 \pm 0.162}$	-0.805 ± 0.104
Kin8nm	$\mathbf{0.084 \pm 0.003}$	0.090 ± 0.003	0.232 ± 0.135	$\mathbf{0.468 \pm 0.090}$
Naval	0.003 ± 0.000	$\mathbf{0.004 \pm 0.000}$	-0.624 ± 0.161	$\mathbf{-0.090 \pm 0.105}$
Combined	$\mathbf{4.028 \pm 0.22}$	4.057 ± 0.218	$\mathbf{-2.335 \pm 0.066}$	-2.354 ± 0.052
Protein	$\mathbf{4.581 \pm 0.026}$	4.600 ± 0.026	-2.526 ± 0.045	$\mathbf{-2.456 \pm 0.017}$
Wine	0.676 ± 0.051	$\mathbf{0.626 \pm 0.045}$	-1.261 ± 0.172	$\mathbf{-0.750 \pm 0.097}$
Yacht	$\mathbf{1.664 \pm 0.607}$	1.861 ± 0.662	$\mathbf{-0.788 \pm 0.511}$	-0.813 ± 0.227
Year	$8.922 \pm \text{NA}$	$\mathbf{8.689 \pm NA}$	$-2.940 \pm \text{NA}$	$\mathbf{-2.872 \pm NA}$