STEIN VARIATIONAL GRADIENT DESCENT ON INFINITE-DIMENSIONAL SPACE AND APPLICATIONS TO STATISTICAL INVERSE PROBLEMS[*]

JUNXIONG JIA[†], PEIJUN LI[‡], AND DEYU MENG[§]

Abstract. In this paper, we propose an infinite-dimensional version of the Stein variational gradient descent (iSVGD) method for solving Bayesian inverse problems. The method can generate approximate samples from posteriors efficiently. Based on the concepts of operator-valued kernels and vector-valued reproducing kernel Hilbert spaces, a rigorous definition is given for the infinite-dimensional objects, e.g., the Stein operator, which are proved to be the limit of finite-dimensional ones. Moreover, a more efficient iSVGD with preconditioning operators is constructed by generalizing the change of variables formula and introducing a regularity parameter. The proposed algorithms are applied to an inverse problem of the steady state Darcy flow equation. Numerical results confirm our theoretical findings and demonstrate the potential applications of the proposed approach in the posterior sampling of large-scale nonlinear statistical inverse problems.

Key words. statistical inverse problems, Bayes' method, variational inference method, Stein variational gradient descent, machine learning

MSC codes. 65L09, 49N45, 62F15

DOI. 10.1137/21M1440773

1. Introduction.

Driven by rapid algorithmic development and a steady increase of computer power, the Bayesian approach has enjoyed great popularity for solving inverse problems over the last decade. By transforming inverse problems into statistical inference problems, the approach provides a general framework to quantify uncertainties [1]. The posterior distribution automatically delivers an estimate of the statistical uncertainty in the reconstruction, and hence suggests "confidence" intervals that allow one to reject or accept scientific hypotheses [44]. It has been widely used in many applications, e.g., artifact detecting in medical imaging [64].

The approach begins with establishing an appropriate Bayes model. When the parameters are in a finite-dimensional space, the finite-dimensional Bayesian method can be employed [56]. A comprehensive account of the finite-dimensional theory can be found in [32]. When the inferred parameters are in the infinite-dimensional space, the problems are more challenging since the Lebesgue measure cannot be defined rigorously in this case [15]. Recently, some attempts have been made to handle the

[*] Received by the editors August 16, 2021; accepted for publication (in revised form) May 10, 2022; published electronically August 18, 2022.
https://doi.org/10.1137/21M1440773

Funding: The work of the first author was supported by the National Key R&D Program of the Ministry of Science and Technology of China grant 2020YFA0713403 and the NSFC grants 11871392, 12090020, 12090021. The work of the second author was partially supported by the NSF grant DMS-1912704. The work of the third author was supported by the NSFC grants 61721002, U1811461, The Major Key Project of PCL (PCL2021A12), and the Macao Science and Technology Development Fund grant 061/2020/A2.

[†] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China (jjx323@xjtu.edu.cn).

[‡] Department of Mathematics, Purdue University, West Lafayette, Indiana, 47907, USA, 47907 (lipeijun@math.purdue.edu).

[§] Corresponding author. School of Mathematics and Statistics and Ministry of Education Key Lab of Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an, 710049, China, Peng Cheng Laboratory, Shenzhen, 518055, China, and Macao Institute of Systems Engineering, Macau University of Science and Technology, Taipa, Macao (dymeng@mail.xjtu.edu.cn).

issue. For example, a general framework was designed for the Bayesian formula and the general theory was applied to inverse problems of fluid mechanic equations [12]. A survey can be found in [53] on the basic framework of the infinite-dimensional Bayes' approach for solving inverse problems. Inverse problems of partial differential equations (PDEs) often involve infinite-dimensional spaces, and the infinite-dimensional Bayes' theory has recently attracted more attention [5, 13, 24, 45, 46].

As pointed out in [1], one of the challenges for the Bayesian approach is how to effectively extract information encoded in the posterior probability measure. To overcome the difficulty, the two main strategies are the point estimate method and the sampling method. The former is to find the maximum a posteriori (MAP) estimate which is equivalent to solving an optimization problem [5, 24]. In some situations, the MAP estimates are more desirable and computationally feasible than the entire posterior distribution [26, 55]. However, the point estimates cannot convey uncertainty information and are usually recognized as an incomplete Bayes' method. The sampling-type methods, such as the well known Markov chain Monte Carlo (MCMC), are often used to extract posterior information. They are well studied in the finite-dimensional setting [35]. Although the MCMC methods are accurate and effective, they are usually not robust under mesh refinement [13]. Multiple dimension-independent MCMC-type algorithms have been proposed [13, 14, 20, 51]. However, these MCMC-type algorithms are computationally too expensive to be adopted in such an application as seismic exploration [21].

The finite-dimensional problems have been extensively studied and many efficient algorithms have been developed to quantify uncertainties effectively. In particular, the variational inference (VI) methods have been broadly investigated in machine learning [3, 43, 62, 63]. Under the mean-field assumption, the linear inverse problems were examined in [30, 29] by using a hierarchical formulation with Gaussian and centered-t noise distribution. The skewed-t noise distribution was considered for a similar setting in [23]. A new type of VI algorithm, called the Stein variational gradient descent (SVGD), was proposed in [39]. The method can achieve a reliable uncertainty estimation by efficiently using an interacting repulsive mechanism. The SVGD has shown to be a fast and flexible method for solving challenging machine learning problems and inverse problems of PDEs [10, 11].

Compared with the finite-dimensional problems, the infinite-dimensional problems are much less studied for VI. When the approximate measures are restricted to be Gaussian, the novel Robbins–Monro algorithm was developed in [45, 46] from a calculus-of-variations viewpoint. It was shown in [54] that the Kullback–Leibler (KL) divergence between the stochastic processes is equal to the supremum of the KL divergence between the measures restricted to finite marginals. Meanwhile, they developed a VI method for functions parameterized by Bayesian neural networks. Under the classical mean-field assumption, a general VI framework defined on separable Hilbert spaces was proposed recently in [28]. A function space particle optimization method including the SVGD was developed in [61] to solve the particle optimization directly in the space of functions. The function space algorithm was also employed to solve computer vision problems, e.g., the context of semantic segmentation and depth estimation [9]. However, the function spaced SVGD assumes that the random functions can be parameterized by a finite number of parameters, e.g., parameterized by some neural networks [61]. Hence, the probability measures on functions are implicitly defined through the probability distributions of a finite number of parameters, instead of the expected infinite-dimensional function space.

This work concerns inverse problems of PDEs imposed on infinite-dimensional function spaces. Motivated by the preconditioned Crank–Nicolson (pCN) algorithm

[13], we aim to construct the SVGD on separable Hilbert spaces with random functions. Throughout, the iSVGD stands for SVGD defined on the infinite-dimensional function space. The goal is to develop algorithms defined on Hilbert spaces and lay a foundation for appropriate discretizations. It contains three contributions:

(1) We investigate the Bayesian formula in infinite-dimensional spaces. The rigorous definition of the SVGD on separable Hilbert spaces is provided, the Stein operator is defined, and the corresponding optimization problem on some Hilbert spaces is considered, and the finite-dimensional problem is proved to converge to the infinite-dimensional counterpart.

(2) By introducing the vector-valued reproducing kernel Hilbert space (RKHS) and operator-valued kernel, we improve the iSVGD with precondition information (e.g., Hessian information operator), which can accelerate the iSVGD algorithm significantly. This is the first work on such an iSVGD algorithm with precondition information.

(3) Explicit numerical strategies are designed by using the finite-element approach. Through theoretical analysis and numerical examples, we demonstrate that the regularity parameter s introduced in the abstract theory (see Assumptions 5 and 7 in section 3.2) should belong to the interval $(0, 0.5)$ and be close to 0.5. The scalability of the algorithm depends only on the scalability of the forward and adjoint PDE solvers. Hence, the algorithm is applicable to solving large-scale inverse problems of PDEs.

The paper is organized as follows. The SVGD in finite-dimensional spaces is introduced in section 2. Section 3 is devoted to the construction of the iSVGD. The basic concepts of operator-valued kernels and Hilbert scales are briefly reviewed; the Stein operator is defined on separable Hilbert spaces; it is shown that the infinite-dimensional version is indeed equivalent to the finite-dimensional version in some limit sense; Based on the Stein operator and the theory of RKHS, the update direction of the iSVGD is derived; in addition, the change of variables is studied and the iSVGD is constructed with preconditioning operators; a preliminary theoretical study is given for the corresponding continuous equations. In section 4, the algorithm is applied to solve an inverse problem governed by the steady state Darcy flow equation. The paper is concluded with some general remarks and directions for future work in section 5.

2. A short review of SVGD.

Let $\mathcal{H}$ be a separable Hilbert space endowed with the Borel σ-algebra $\mathcal{B}(\mathcal{H})$. Denote by $\mathcal{G}$, u, and $\boldsymbol{d}$ the solution operator of some PDE, the model parameter, and the observation, respectively. We assume that $u \in \mathcal{H}$ and $\boldsymbol{d} \in \mathbb{R}^{N_d}$ with N_d being a positive integer. The observation $\boldsymbol{d}$ is related to $\mathcal{G}(u)$ and the random noise $\boldsymbol{\epsilon}$ through some functions [32], e.g., the additive noise model or the multiplicative noise model. We refer to section 4 for a specific example.

For statistical inverse problems, it is usually required to find a probability measure $\mu^{\boldsymbol{d}}$ on $\mathcal{H}$, which is known as the posterior probability measure and is specified by its density with respect to a prior probability measure μ_0. The Bayesian formula on a Hilbert space is defined by

$$\frac{d\mu^{\boldsymbol{d}}}{d\mu_0}(u) = \frac{1}{Z_{\boldsymbol{d}}} \exp\Big(- \Phi(u; \boldsymbol{d})\Big), \tag{1}$$

where $\Phi \in C(\mathcal{H} \times \mathbb{R}^{N_d}; \mathbb{R})$ and $\exp(-\Phi(u; \boldsymbol{d}))$ is integrable with respect to μ_0. The constant $Z_{\boldsymbol{d}}$ is chosen to ensure that $\mu^{\boldsymbol{d}}$ is indeed a probability measure. The prior measure $\mu_0 := \mathcal{N}(0, \mathcal{C}_0)$ is assumed to be a Gaussian measure defined on $\mathcal{H}$ with $\mathcal{C}_0$

being a self-adjoint, positive definite, and trace class operator. Let $(\lambda_k, \varepsilon_k)_{k=1}^{\infty}$ be the eigensystem of $\mathcal{C}_0$ satisfying $\mathcal{C}_0\varepsilon_k = \lambda_k^2\varepsilon_k$. Denote by P^N and Q^N the orthogonal projections of $\mathcal{H}$ onto $X^N := \text{span}\{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N\}$ and $X^\perp := \text{span}\{\varepsilon_{N+1}, \varepsilon_{N+2}, \dots\}$, respectively. Clearly, we have $Q^N = \text{Id} - P^N$. Let $u^N := P^N u \in X^N$ and $u^\perp := Q^N u \in X^\perp$. Define $\mathcal{C}_0^N = P^N\mathcal{C}_0P^N$ and let $\mu_0^N = \mathcal{N}(0, \mathcal{C}_0^N)$ be a finite-dimensional Gaussian measure defined on X^N. Then an approximate measure $\mu^{\boldsymbol{d}N}$ on X^N can be defined by

$$\frac{d\mu^{\boldsymbol{d}N}}{d\mu_0^N}(u^N) = \frac{1}{Z_{\boldsymbol{d}}^N}\exp\Big(-\Phi(u^N;\boldsymbol{d})\Big), \tag{2}$$

where

$$Z_{\boldsymbol{d}}^N = \int_{X^N}\exp\Big(-\Phi(u^N;\boldsymbol{d})\Big)\mu_0^N(du^N).$$

Some more properties of the above approximate measure can be found in [16, subsection 5.6]. The probability measure $\mu^{\boldsymbol{d}N}$ can be written as the pushforward of the posterior measure $\mu^{\boldsymbol{d}}$ on $\mathbb{R}^N$, i.e., $\mu^{\boldsymbol{d}N} = P_{\#}^N\mu^{\boldsymbol{d}} := \mu^{\boldsymbol{d}} \circ (P^N)^{-1}$. Hence the measure $\mu^{\boldsymbol{d}N}$ has a Lebesgue density denoted by $p^{\boldsymbol{d}N}$ with the following form,

$$p^{\boldsymbol{d}N}(u^N) \propto \exp\Big(-\Phi(u^N;\boldsymbol{d}) - \frac{1}{2}\|u^N\|_{\mathcal{C}_0^N}^2\Big), \tag{3}$$

where $\|\cdot\|_{\mathcal{C}_0^N}$ represents $\|(\mathcal{C}_0^N)^{-1/2}\cdot\|_{\ell^2}$ with $\|\cdot\|_{\ell^2}$ standing for the usual ℓ^2-norm. Obviously, the target distribution $\mu^{\boldsymbol{d}N}$ is the solution to the optimization problem defined on the set $\mathcal{P}_2(\mathbb{R}^N)$ of probability measures ν such that $\int\|u^N\|^2d\nu^N(u^N) < \infty$ by

$$\min_{\nu^N\in\mathcal{P}_2(\mathbb{R}^N)} \text{KL}(\nu^N||\mu^{\boldsymbol{d}N}), \tag{4}$$

where KL denotes the KL divergence.

Now, we present the SVGD algorithm. Denote $\text{KL}(\cdot||\mu^{\boldsymbol{d}N}) : \mathcal{P}_2(\mathbb{R}^N) \to [0, +\infty)$ as the functional $\nu^N \mapsto \text{KL}(\nu^N||\mu^{\boldsymbol{d}N})$. In order to obtain samples from $\mu^{\boldsymbol{d}N}$, the SVGD applies a gradient descent-like algorithm to the functional $\text{KL}(\cdot||\mu^{\boldsymbol{d}N})$. The standard gradient descent algorithm in the Wasserstein space applied to $\text{KL}(\cdot||\mu^{\boldsymbol{d}N})$, at each iteration $\ell \geq 0$, is

$$\nu_{\ell+1}^N = \left(\text{Id} - \epsilon\nabla\log\left(\frac{d\nu_\ell^N}{d\mu^{\boldsymbol{d}N}}\right)\right)_{\#}\nu_\ell^N, \tag{5}$$

where $\epsilon > 0$ is the step size. This corresponds to a forward Euler discretization of the gradient flow of $\text{KL}(\cdot||\mu^{\boldsymbol{d}N})$ with respect to Stein geometry [18]. Instead of the Wasserstein gradient $\nabla\log\big(d\nu_\ell^N/d\mu^{\boldsymbol{d}N}\big)$ used in (5), the SVGD uses $P_{\nu_\ell^N}\nabla\log\big(d\nu_\ell^N/d\mu^{\boldsymbol{d}N}\big)$ to generate the following iteration,

$$\nu_{\ell+1}^N = \left(\text{Id} - \epsilon P_{\nu_\ell^N}\nabla\log\left(\frac{d\nu_\ell^N}{d\mu^{\boldsymbol{d}N}}\right)\right)_{\#}\nu_\ell^N, \tag{6}$$

where $P_{\nu_\ell^N}$ is the same as that in subsection 3.1 of [33]. Let $\mathcal{H}_K^N$ be an N-dimensional RKHS [52] with the kernel function $K : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$. To define $P_{\nu_\ell^N}$ rigorously, it is necessary to introduce the kernel integral operator based on the kernel function K, which will not be used in the rest of the paper. Hence, we omit it and refer to [33] for the details. The reason for introducing the operator $P_{\nu_\ell^N}$ is that we have

$$P_{\nu_\ell^N}\nabla\log\left(\frac{d\nu_\ell^N}{d\mu^{\boldsymbol{d}N}}\right)(\cdot) = -\mathbb{E}_{u^N\sim\nu_\ell^N}\left[K(u^N,\cdot)\nabla_{u^N}\log p^{\boldsymbol{d}N}(u^N) + \nabla_{u^N}K(u^N,\cdot)\right] \tag{7}$$

under some mild conditions. For every $\ell \geq 0$, let $u^{N,\ell}$ be distributed according to ν_ℓ^N. Using (6)–(7), we obtain a particle update scheme

$$u^{N,\ell+1} = u^{N,\ell} + \epsilon\phi_\ell^N(u^{N,\ell}), \tag{8}$$

where

$$\phi_\ell^{N*}(\cdot) = \mathbb{E}_{u^N\sim q_\ell^N}\Big[K(u^N,\cdot)\nabla_{u^N}\log p^{\boldsymbol{d}N}(u^N) + \nabla_{u^N}K(u^N,\cdot)\Big]. \tag{9}$$

The basic SVGD algorithm is given in Algorithm 1. Inspired by applications in machine learning, the SVGD-type algorithms have been widely studied over the last few years [17, 18, 33, 38, 39, 42].

Algorithm 1. Finite-dimensional SVGD.

```
Input: A target probability measure with density function $p^{\boldsymbol{d}N}(u^N)$ and a set of particles $\{u_i^{N,0}\}_{i=1}^m$.
Output: A set of particles $\{u_i^N\}_{i=1}^m$ that approximates the target probability measure.
for iteration $\ell$ do
    $u_i^{N,\ell+1} \longleftarrow u_i^{N,\ell} + \epsilon_\ell\phi^*(u_i^{N,\ell})$,
    where
    $\phi^*(u^N) = \frac{1}{m}\sum_{j=1}^{m}\Big[K(u_j^{N,\ell},u^N)\nabla_{u_j^{N,\ell}}\log p^{\boldsymbol{d}N}(u_j^{N,\ell}) + \nabla_{u_j^{N,\ell}}K(u_j^{N,\ell},u^N)\Big]$,
    and $\epsilon_\ell$ is the step size at the $\ell$th iteration.
end for
```

3. SVGD on separable Hilbert spaces.

This section is devoted to the construction of iSVGD and the preconditioning operators. The corresponding continuity equations are provided for a preliminary theoretical study of the method.

3.1. Hilbert scale and vector-valued RKHS.

For constructing iSVGD, we need to characterize the smoothness of functions that belong to some infinite-dimensional spaces. The Sobolev spaces are usually employed to characterize the smoothness of functions. However, for presenting a general theory, we introduce the Hilbert scales defined by the prior covariance operator [19]. The reason is that different covariance operators employed in practical problems lead to the same form of Hilbert scales. However, they are related to different Sobolev spaces. Hence, the same form of the general theory can be flexibly adapted to different practical problems.

Let $\mathcal{C}_0 : \mathcal{H} \to \mathcal{H}$ be the covariance operator introduced in section 2. Denote by $\mathcal{D}(\mathcal{C}_0)$ and $\mathcal{R}(\mathcal{C}_0)$ the domain and range of $\mathcal{C}_0$, respectively. Let $\mathcal{H} = \overline{\mathcal{R}(\mathcal{C}_0)} \oplus \mathcal{R}(\mathcal{C}_0)^\perp = \overline{\mathcal{R}(\mathcal{C}_0)}$ (the closure of $\mathcal{R}(\mathcal{C}_0)$). It is clear to note that $\mathcal{C}_0^{-1}$ is a densely defined, unbounded, symmetric and positive-definite operator in $\mathcal{H}$. Let $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$ be the inner product and norm defined on the Hilbert space $\mathcal{H}$, respectively. Define the Hilbert scales $(\mathcal{H}^t)_{t\in\mathbb{R}}$ with $\mathcal{H}^t := \overline{\mathcal{S}_f}^{\|\cdot\|_{\mathcal{H}^t}}$, where

$$\mathcal{S}_f := \bigcap_{n=0}^{\infty} \mathcal{D}(\mathcal{C}_0^{-n}), \quad \langle u,v\rangle_{\mathcal{H}^t} := \langle \mathcal{C}_0^{-t/2}u, \mathcal{C}_0^{-t/2}v\rangle_{\mathcal{H}}, \quad \|u\|_{\mathcal{H}^t} := \left\|\mathcal{C}_0^{-t/2}u\right\|_{\mathcal{H}}.$$

The norms defined above possess the following properties (cf. [19, Proposition 8.19]).

LEMMA 1. *Let $(\mathcal{H}^t)_{t\in\mathbb{R}}$ be the Hilbert scale induced by the operator $\mathcal{C}_0$ given above. Then the following assertions hold:*

1. *Let $-\infty < s < t < \infty$. Then the space $\mathcal{H}^t$ is densely and continuously embedded into $\mathcal{H}^s$.*
2. *If $t \geq 0$, then $\mathcal{H}^t = \mathcal{D}(\mathcal{C}_0^{-t/2})$, and $\mathcal{H}^{-t}$ is the dual space of $\mathcal{H}^t$.*
3. *Let $-\infty < q < r < s < \infty$ then the interpolation inequality $\|u\|_{\mathcal{H}^r} \leq \|u\|_{\mathcal{H}^q}^{\frac{s-r}{s-q}} \|u\|_{\mathcal{H}^s}^{\frac{r-q}{s-q}}$ holds when $u \in \mathcal{H}^s$.*

Now, we introduce some basic notations of vector-valued RKHS. The following definition concerns the Hilbert space adjoint opertor [50].

DEFINITION 2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces, and T be a bounded linear operator from $\mathcal{X}$ to $\mathcal{Y}$. The* **Banach space adjoint** *of T, denoted by T', is the bounded linear operator from $\mathcal{Y}^*$ to $\mathcal{X}^*$ and is defined by $(T'\ell)(u) = \ell(Tu)$ for all $\ell \in \mathcal{Y}^*$, $u \in \mathcal{X}$. Let $\mathcal{X}$ and $\mathcal{Y}$ be Hilbert spaces, and $C_1 : \mathcal{X} \to \mathcal{X}^*$ be the map that assigns to each $u \in \mathcal{X}$, the bounded linear functional $\langle u,\cdot\rangle_{\mathcal{X}}$ in $\mathcal{X}^*$. Let $C_2 : \mathcal{Y} \to \mathcal{Y}^*$ be defined similarly as C_1. Then the* **Hilbert space adjoint** *of T is a map $T^* : \mathcal{Y} \to \mathcal{X}$ given by $T^* = C_1^{-1}T'C_2$.*

Next, we introduce operator-valued positive-definite kernels, which constitute the framework for specifying vector-valued RKHS. Following Kadri et al. [31] to avoid topological and measurability issues, we focus on separable Hilbert spaces with reproducing operator-valued kernels whose elements are continuous functions. Denote by $\mathcal{X}$ and $\mathcal{Y}$ the separable Hilbert spaces and by $\mathcal{L}(\mathcal{X},\mathcal{Y})$ the set of bounded linear operators from $\mathcal{X}$ to $\mathcal{Y}$. When $\mathcal{X} = \mathcal{Y}$, we write $\mathcal{L}(\mathcal{Y},\mathcal{Y})$ briefly as $\mathcal{L}(\mathcal{Y})$.

DEFINITION 3 (operator-valued kernels). *An $\mathcal{L}(\mathcal{Y})$-valued kernel $\boldsymbol{K}$ on $\mathcal{X}\times\mathcal{X}$ is an operator $\boldsymbol{K}(\cdot,\cdot) : \mathcal{X}\times\mathcal{X} \to \mathcal{L}(\mathcal{Y})$.*

1. *$\boldsymbol{K}$ is Hermitian if $\forall\, u,v \in \mathcal{X}$, $\boldsymbol{K}(u,v) = \boldsymbol{K}(v,u)^*$;*
2. *$\boldsymbol{K}$ is nonnegative on $\mathcal{X}$ if it is Hermitian and for every natural number r and all $\{(u_i,v_i)_{i=1,\ldots,r}\} \in \mathcal{X}\times\mathcal{Y}$, the matrix with ijth entry $\langle \boldsymbol{K}(u_i,u_j)v_i, v_j\rangle_{\mathcal{Y}}$ is nonnegative (positive definite).*

DEFINITION 4 (vector-valued RKHS). *Let $\mathcal{X}$ and $\mathcal{Y}$ be separable Hilbert spaces. A Hilbert space $\mathcal{F}$ of operators from $\mathcal{X}$ to $\mathcal{Y}$ is called an RKHS if there is a nonnegative $\mathcal{L}(\mathcal{Y})$-valued kernel $\boldsymbol{K}$ on $\mathcal{X}\times\mathcal{X}$ such that*

1. *the operator $v \longmapsto \boldsymbol{K}(u,v)g$ belongs to $\mathcal{F}$ for all $v,u \in \mathcal{X}$ and $g \in \mathcal{Y}$;*
2. *for every $f \in \mathcal{F}$, $u \in \mathcal{X}$, and $g \in \mathcal{Y}$, we have $\langle f(u), g\rangle_{\mathcal{Y}} = \langle f(\cdot), \boldsymbol{K}(u,\cdot)g\rangle_{\mathcal{F}}$.*

Throughout the paper, we assume that the kernel $\boldsymbol{K}$ is *locally bounded* and *separately continuous*, which guarantee that $\mathcal{F}$ is a subspace of $C(\mathcal{X},\mathcal{Y})$ (the vector space of continuous operators from $\mathcal{X}$ to $\mathcal{Y}$). If the kernel $\boldsymbol{K}$ is nice enough [7, 8], then it is the reproducing kernel of some Hilbert space $\mathcal{F}$.

Since the kernel is an important part of the SVGD, we provide some intuitive ideas about the operator-valued kernel. Let $u,v \in \mathcal{H}$ and $h > 0$ be a positive constant.

To construct the infinite-dimensional SVGD, we may introduce a scalar-valued kernel $K(u,v) := \exp\left(-\frac{1}{h}\|u-v\|_{\mathcal{H}}^2\right)$ and consider the operator-valued kernel

$$\boldsymbol{K}(u,v) = K(u,v)\mathrm{Id}. \tag{10}$$

For example, we can take $\mathcal{H} = L^2(\Omega)$ with Ω being a bounded open domain and have

$$\|u-v\|_{\mathcal{H}}^2 = \int_{\Omega} |u(x)-v(x)|^2 dx. \tag{11}$$

However, for solving inverse problems of PDEs, it is useful to introduce some preconditioning operators which require us to consider operator-valued kernels. Here, we illustrate this by a simple example. Let the prior measure $\mu_0 = \mathcal{N}(0, (\mathrm{Id}-\Delta)^{-2})$, where Δ is the Dirichlet Laplace operator and $\mathcal{H} = L^2(\Omega)$. Intuitively we have $\mathcal{H}^1 \approx H^2(\Omega)$, where $H^2(\Omega)$ is the usual Sobolev space. By the theory of Gaussian measures [48], we approximately have $\mu_0(H^2(\Omega)) = 0$ (not rigorously correct). Inspired by the pCN algorithm [13], we may choose the preconditioning operator $T = \mathrm{Id} - \Delta$. If we choose the Gaussian kernel as (10), then the transformed kernel function becomes

$$\boldsymbol{K}(u,v) = \exp\left(-\frac{1}{h}\|T(u-v)\|_{L^2}^2\right) T^{-1}(T^{-1})^*, \tag{12}$$

which is approximately equal to

$$\boldsymbol{K}(u,v) \approx \exp\left(-\frac{1}{h}\|u-v\|_{H^2}^2\right) (\mathrm{Id}-\Delta)^{-2}. \tag{13}$$

Obviously, the kernel function equals zero when $u-v$ does not belong to $H^2(\Omega)$, i.e., $\|u-v\|_{H^2} < \infty$ when $u-v \in H^2(\Omega)$. Hence, the kernel function takes nonzero values and the algorithms can work only if the differences of any two particles reside in a measure zero set. In our opinion, this restriction seems too strong in the infinite-dimensional setting to make the particles overconcentrated (see our numerical example in section 4 to demonstrate this in detail).

Based on the above discussion, we may introduce a parameter s and have an approximate transformed kernel

$$\boldsymbol{K}(u,v) \approx \exp\left(-\frac{1}{h}\|u-v\|_{H^{2-2s}}^2\right) (\mathrm{Id}-\Delta)^{-2}. \tag{14}$$

However, to achieve this, we should not choose the original kernel (the kernel is not transformed by the operator T) to be the usual scalar-valued kernel. The original kernel may be chosen as $\boldsymbol{K}_0(u,v) = K_0(u,v)(\mathrm{Id}-\Delta)^{-2s}$, where $K_0(u,v) := e^{-\frac{1}{h}\|u-v\|_{L^2}}$ with $h>0$ being a positive constant. In this setting, the preconditioning operator can be chosen as $T := (\mathrm{Id}-\Delta)^{1-s}$. These intuitive ideas indicate that it is necessary to construct the infinite-dimensional SVGD based on the more involved operator-valued kernel theory.

3.2. iSVGD.

In this subsection, we present an infinite-dimensional version of the SVGD, i.e., iSVGD. For a function u, denote by D_u and D_{u_k} the Fréchet derivative and the directional derivative in the kth direction, respectively. For simplicity of notation, we shall use D and D_k instead of D_u and D_{u_k}, and write $\Phi(u;\boldsymbol{d})$ as $\Phi(u)$. Let

$$V(u) = \Phi(u) + \frac{1}{2}\|u\|_{\mathcal{H}^1}^2, \tag{15}$$

where the potential functional Φ is required to satisfy the following assumptions.

Assumption 5. Let $\mathcal{X}$ and $\mathcal{H}$ be two separable Hilbert spaces. For $s \in [0, 1]$, we assume $\mathcal{H}^{1-s} \subset \mathcal{X} \subset \mathcal{H}$. Let $M_1 \in \mathbb{R}^+$ be a positive constant. For each $u \in \mathcal{X} \subset \mathcal{H}$, we introduce $D\Phi : \mathcal{X} \to \mathcal{X}^*$ and $D^2\Phi : \mathcal{X} \to \mathcal{L}(\mathcal{X}, \mathcal{X}^*)$, then the functional $\Phi : \mathcal{X} \to \mathbb{R}$ satisfies

$$\begin{aligned} -M_1 \leq \Phi(u) &\leq M_2(\|u\|_{\mathcal{X}}), \\ \|D\Phi(u)\|_{\mathcal{X}^*} &\leq M_3(\|u\|_{\mathcal{X}}), \\ \|D^2\Phi(u)\|_{\mathcal{L}(\mathcal{X},\mathcal{X}^*)} &\leq M_4(\|u\|_{\mathcal{X}}), \end{aligned}$$

where $M_2(\cdot)$, $M_3(\cdot)$, and $M_4(\cdot)$ are some monotonic nondecreasing functions.

The above assumption is a local version of [16, Assumption 4], which can be verified for many problems, e.g., the Darcy flow model (Theorem 17 in section 4). We now optimize ϕ in the unit ball of a general vector-valued RKHS $\mathcal{H}_{\boldsymbol{K}}$ with an operator-valued kernel $\boldsymbol{K}(u, u') \in \mathcal{L}(\mathcal{Y})$:

$$(16) \qquad \phi^*_{\boldsymbol{K}} = \underset{\phi \in \mathcal{H}_{\boldsymbol{K}}}{\arg\max} \left\{ \mathbb{E}_{u\sim\mu}[\mathcal{S}\phi(u)] \text{ s.t. } \|\phi\|_{\mathcal{H}_K} \leq 1 \text{ and } D\phi : \mathcal{X} \to \mathcal{L}_1(\mathcal{X}, \mathcal{Y}) \right\},$$

where $\mathcal{S}$ is the generalized Stein operator defined formally as follows,

$$(17) \qquad \mathcal{S}\phi(u) = -\langle DV(u), \phi(u)\rangle_{\mathcal{Y}} + \sum_{k=1}^{\infty} D_k \langle \phi(u), e_k\rangle_{\mathcal{Y}},$$

and $\mathcal{L}_1(\mathcal{X}, \mathcal{Y})$ denotes the set of all trace class operators from $\mathcal{X}$ to $\mathcal{Y}$. For the convergence of the infinite sum, we illustrate it in Theorem 9. Here, $\{e_k\}_{k=1}^{\infty}$ stands for an orthonormal basis of space $\mathcal{Y}$ and μ is a probability measure defined on $\mathcal{H}$. Moreover, we assume that $\phi : \mathcal{X} \to \mathcal{Y}$ is Fréchet differentiable, and the derivative is continuous to ensure the validity of (16).

Remark 6. In the finite-dimensional case, the operator $D\phi(u)$ naturally belongs to $\mathcal{L}_1(\mathcal{X}, \mathcal{Y})$ (cf. [15, Appendix C]).

The following assumption is also needed for the operator-valued kernels, which include many useful kernels, e.g., the radial basis function (RBF) kernel.

Assumption 7. Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{H}$ be three separable Hilbert spaces. For $s \in [0, 1]$, we assume that $\mathcal{H}^{-s-1} \subset \mathcal{Y}$ and

$$(18) \qquad \sup_{u \in \mathcal{X}} \|\boldsymbol{K}(u, u)\|_{\mathcal{L}(\mathcal{Y})} < \infty.$$

Remark 8. We mention that condition (18) holds for the bounded scalar-valued kernel functionals since a scalar-valued kernel functional can be seen as a scalar-valued kernel functional composite with an identity operator as demonstrated in (10).

To illustrate (16) and (17), we prove Theorem 9. For each particle u, we assume that $u \in \mathcal{H}^{1-s}$, which is based on the following two considerations:

- The SVGD with one particle is an optimization algorithm for finding the MAP estimate. The MAP estimate belongs to the separable Hilbert space $\mathcal{H}^1$.
- For the prior probability measure, the space $\mathcal{H}^1$ has zero measure [15]. Intuitively, if all particles belong to $\mathcal{H}^1$, the particles tend to concentrate around a small set that leads to unreliable estimates of statistical quantities. Hence, we may assume that the particles belong to a larger space containing $\mathcal{H}^1$.

THEOREM 9. *The generalized Stein operator* (17) *defined on $\mathcal{Y}$ can be obtained by taking $N \to \infty$ in the following finite-dimensional Stein operator,*

$$\mathcal{S}^N \phi^N(u^N) = -\langle DV(u^N), \phi^N(u^N)\rangle_{\mathcal{Y}} + \sum_{k=1}^{N} D_k \langle \phi^N(u^N), e_k\rangle_{\mathcal{Y}}, \tag{19}$$

where $\phi^N := P^N \circ \phi$.

Proof. By straightforward calculations, we have

$$\begin{aligned} \mathcal{S}\phi(u) - \mathcal{S}^N\phi^N(u^N) &= -\Big(\langle DV(u), \phi(u)\rangle_{\mathcal{Y}} - \langle DV(u^N), \phi^N(u^N)\rangle_{\mathcal{Y}}\Big) \\ &\quad + \left(\sum_{k=1}^{\infty} D_k\langle \phi(u), e_k\rangle_{\mathcal{Y}} - \sum_{k=1}^{N} D_k\langle \phi^N(u^N), e_k\rangle_{\mathcal{Y}}\right) \\ &= -\,\mathrm{I} + \mathrm{II}. \end{aligned} \tag{20}$$

For term I, we have

$$\begin{aligned} \mathrm{I} &= \langle D(V(u) - V(u^N)), \phi^N(u^N)\rangle_{\mathcal{Y}} + \langle DV(u), \phi(u) - \phi^N(u^N)\rangle_{\mathcal{Y}} \\ &= \mathrm{I}_1(N) + \mathrm{I}_2(N). \end{aligned} \tag{21}$$

For term $\mathrm{I}_1(N)$, we find that

$$\mathrm{I}_1(N) = \langle D(\Phi(u) - \Phi(u^N)), \phi^N(u^N)\rangle_{\mathcal{Y}} + \langle \mathcal{C}_0^{-1/2}(u - u^N), \mathcal{C}_0^{-1/2}\phi^N(u^N)\rangle_{\mathcal{Y}}, \tag{22}$$

where the second term on the right-hand side is understood as the white noise mapping [48]. According to Assumptions 5 and 7, we know that

$$\begin{aligned} \lim_{N\to\infty} \|D(\Phi(u) - \Phi(u^N))\|_{\mathcal{Y}} &\le \lim_{N\to\infty} C\|D(\Phi(u) - \Phi(u^N))\|_{\mathcal{H}^{-1-s}} \\ &\le \lim_{N\to\infty} C\|D(\Phi(u) - \Phi(u^N))\|_{\mathcal{H}^{-1+s}} \\ &\le \lim_{N\to\infty} CM_4(2\|u\|_{\mathcal{X}})\|u - u^N\|_{\mathcal{H}^{1-s}} = 0, \end{aligned} \tag{23}$$

where C is a generic constant that can be different from line to line. Hence, we obtain

$$\lim_{N\to\infty} \langle D(\Phi(u) - \Phi(u^N)), \phi^N(u^N)\rangle_{\mathcal{Y}} = 0. \tag{24}$$

Taking $u_m \in \mathcal{H}^2$ such that $u_m \to u$ in $\mathcal{H}^{1-s}$, we have

$$\begin{aligned} \langle \mathcal{C}_0^{-1/2}(u - u^N), \mathcal{C}_0^{-1/2}\phi^N(u^N)\rangle_{\mathcal{Y}} &= \lim_{m\to\infty} \langle \mathcal{C}_0^{-1/2}(u_m - u_m^N), \mathcal{C}_0^{-1/2}\phi^N(u^N)\rangle_{\mathcal{Y}} \\ &= \lim_{m\to\infty} \langle P^N\mathcal{C}_0^{-1}(u_m - u_m^N), \phi(u^N)\rangle_{\mathcal{Y}} \\ &= \lim_{m\to\infty} \langle \phi(\cdot), \boldsymbol{K}(u^N, \cdot)P^N\mathcal{C}_0^{-1}(u_m - u_m^N)\rangle_{\mathcal{H}_{\boldsymbol{K}}}. \end{aligned}$$

As for the last term in the above equality, we have the following estimates:

$$\begin{aligned} &\langle \phi(\cdot), \boldsymbol{K}(u^N, \cdot)P^N\mathcal{C}_0^{-1}(u_m - u_m^N)\rangle_{\mathcal{H}_{\boldsymbol{K}}} \le \\ &\qquad \langle \phi(\cdot), \phi(\cdot)\rangle_{\mathcal{H}_{\boldsymbol{K}}} \langle \boldsymbol{K}(u^N, \cdot)P^N\mathcal{C}_0^{-1}(u_m - u_m^N), \boldsymbol{K}(u^N, \cdot)P^N\mathcal{C}_0^{-1}(u_m - u_m^N)\rangle_{\mathcal{H}_{\boldsymbol{K}}} \\ &\quad \le \langle \phi(\cdot), \phi(\cdot)\rangle_{\mathcal{H}_{\boldsymbol{K}}} \langle P^N\boldsymbol{K}(u^N, u^N)P^N\mathcal{C}_0^{-1}(u_m - u_m^N), \mathcal{C}_0^{-1}(u_m - u_m^N)\rangle_{\mathcal{Y}} \\ &\quad \le C\langle \phi(\cdot), \phi(\cdot)\rangle_{\mathcal{H}_{\boldsymbol{K}}} \|\mathcal{C}_0^{-1}(u_m - u_m^N)\|_{\mathcal{Y}}^2 \\ &\quad \le C\langle \phi(\cdot), \phi(\cdot)\rangle_{\mathcal{H}_{\boldsymbol{K}}} \|\mathcal{C}_0^{-\frac{1-s}{2}}(u_m - u_m^N)\|_{\mathcal{H}}^2. \end{aligned}$$

Replacing $u_m - u_m^N$ by $(u_m - u_m^N) - (u - u^N)$, we deduce

$$\begin{aligned}\langle\mathcal{C}_0^{-1/2}(u-u^N), \mathcal{C}_0^{-1/2}\phi^N(u^N)\rangle_{\mathcal{Y}} &= \lim_{m\to\infty}\langle\phi(\cdot), \boldsymbol{K}(u^N,\cdot)P^N\mathcal{C}_0^{-1}(u_m-u_m^N)\rangle_{\mathcal{H}_{\boldsymbol{K}}} \\ &= \langle\phi(\cdot), \boldsymbol{K}(u^N,\cdot)P^N\mathcal{C}_0^{-1}(u-u^N)\rangle_{\mathcal{H}_{\boldsymbol{K}}}.\end{aligned} \tag{25}$$

Hence, we obtain

$$\begin{aligned}&\lim_{N\to\infty}\langle\mathcal{C}_0^{-1/2}(u-u^N), \mathcal{C}_0^{-1/2}\phi^N(u^N)\rangle_{\mathcal{Y}} \\ &\quad = \lim_{N\to\infty}\langle\phi(\cdot), \boldsymbol{K}(u^N,\cdot)P^N\mathcal{C}_0^{-1}(u-u^N)\rangle_{\mathcal{H}_{\boldsymbol{K}}} \\ &\quad \leq \lim_{N\to\infty}\langle\phi(\cdot),\phi(\cdot)\rangle_{\mathcal{H}_{\boldsymbol{K}}}\langle P^N\boldsymbol{K}(u^N,u^N)P^N\mathcal{C}_0^{-1}(u-u^N), \mathcal{C}_0^{-1}(u-u^N)\rangle_{\mathcal{Y}} \\ &\quad \leq C\langle\phi(\cdot),\phi(\cdot)\rangle_{\mathcal{H}_{\boldsymbol{K}}}\lim_{N\to\infty}\|\mathcal{C}_0^{-\frac{1-s}{2}}(u-u^N)\|_{\mathcal{H}}^2 = 0.\end{aligned} \tag{26}$$

Plugging (24) and (26) into (22), we arrive at $\lim_{N\to\infty} I_1(N) = 0$. For term $\mathrm{I}_2(N)$, it can be decomposed as follows:

$$\mathrm{I}_2(N) = \langle D\Phi(u), \phi(u) - \phi^N(u^N)\rangle_{\mathcal{Y}} + \langle\mathcal{C}_0^{-1/2}u, \mathcal{C}_0^{-1/2}(\phi(u)-\phi^N(u^N))\rangle_{\mathcal{Y}}. \tag{27}$$

It follows from the continuity of ϕ that we have $\lim_{N\to\infty}\langle D\Phi(u), \phi(u) - \phi^N(u^N)\rangle_{\mathcal{Y}} = 0$. Using similar estimates as those for deriving (25), we obtain

$$\begin{aligned}&\langle\mathcal{C}_0^{-1/2}u, \mathcal{C}_0^{-1/2}(\phi(u)-\phi^N(u^N))\rangle_{\mathcal{Y}} \\ &\qquad = \langle\phi(\cdot), \boldsymbol{K}(u,\cdot)\mathcal{C}_0^{-1}u\rangle_{\mathcal{H}_{\boldsymbol{K}}} - \langle\phi(\cdot), \boldsymbol{K}(u^N,\cdot)P^N\mathcal{C}_0^{-1}u\rangle_{\mathcal{H}_{\boldsymbol{K}}}.\end{aligned} \tag{28}$$

By the continuity of $\boldsymbol{K}(\cdot,\cdot)$, we obtain

$$\lim_{N\to\infty}\langle\mathcal{C}_0^{-1/2}u, \mathcal{C}_0^{-1/2}(\phi(u)-\phi^N(u^N))\rangle_{\mathcal{Y}} = 0. \tag{29}$$

Now, we conclude that $\lim_{N\to\infty}\mathrm{I}_2(N) = 0$. For term II, we have

$$\mathrm{II} = \sum_{k=1}^{N} D_k\langle\phi(u)-\phi(u^N), e_k\rangle_{\mathcal{Y}} + \sum_{k=N+1}^{\infty} D_k\langle\phi(u), e_k\rangle_{\mathcal{Y}}. \tag{30}$$

Let $\{\varphi_k\}_{k=1}^{\infty}$ be an orthonormal basis in $\mathcal{X}$, and then we have

$$\sum_{k=N+1}^{\infty} D_k\langle\phi(u), e_k\rangle_{\mathcal{Y}} = \sum_{k=N+1}^{\infty}\langle D\phi(u)\varphi_k, e_k\rangle_{\mathcal{Y}} \to 0 \quad \text{as } N\to\infty, \tag{31}$$

where we use the condition $D\phi(u) \in \mathcal{L}_1(\mathcal{X},\mathcal{Y})$. For the first term on the right-hand side of (30), we find that

$$\sum_{k=1}^{N} D_k\langle\phi(u)-\phi(u^N), e_k\rangle_{\mathcal{Y}} = \sum_{k=1}^{N}\langle(D\phi(u)-D\phi(u^N))\varphi_k, e_k\rangle_{\mathcal{Y}}. \tag{32}$$

Due to the continuity of the Fréchet derivative of ϕ, we know that the above summation goes to 0 as $N\to\infty$. Combining the estimates of I and II, we complete the proof. □

The following theorem gives explicitly the iSVGD update directions that are essential for the construction of iSVGD.

THEOREM 10. *Let $\boldsymbol{K}(\cdot,\cdot):\mathcal{X}^2\to\mathcal{L}(\mathcal{Y})$ be a positive-definite kernel that is Fréchet differentiable on both variables. In addition, we assume that*

$$\mathbb{E}_{u\sim\mu}\left[D_{u'}\boldsymbol{K}(u,u')\mathcal{C}_0^{-1/2}g+\sum_{k=1}^{\infty}D_{u_k}D_{u'}\boldsymbol{K}(u,u')e_k\right] \tag{33}$$

belongs to $\mathcal{L}_1(\mathcal{X},\mathcal{Y})$ for each $u'\in\mathcal{X}$ and $g\in\mathcal{H}^{-s}$. Then, the optimal $\phi^*_{\boldsymbol{K}}$ in (16) *is*

$$\phi^*_{\boldsymbol{K}}(\cdot)\propto\mathbb{E}_{u\sim\mu}\left[\boldsymbol{K}(u,\cdot)(-D\Phi(u)-\mathcal{C}_0^{-1}u)+\sum_{k=1}^{\infty}D_{u_k}\boldsymbol{K}(u,\cdot)e_k\right], \tag{34}$$

where $\{e_k\}_{k=1}^{\infty}$ is an orthonormal basis of $\mathcal{Y}$ and the term $\boldsymbol{K}(u,\cdot)\mathcal{C}_0^{-1}u$ is understood in the following limiting sense:

$$\boldsymbol{K}(u,\cdot)\mathcal{C}_0^{-1}u:=\lim_{m\to\infty}\boldsymbol{K}(u,\cdot)\mathcal{C}_0^{-1}u_m. \tag{35}$$

Here the limit is taken in $\mathcal{H}_{\boldsymbol{K}}$ and $\{u_m\}_{m=1}^{\infty}\subset\mathcal{H}^2$ such that $\|\mathcal{C}_0^{-\frac{1-s}{2}}(u_m-u)\|_{\mathcal{H}}\to 0$ as $m\to\infty$.

Proof. First, by taking $\phi(u)$ as an element in $\mathcal{H}_{\boldsymbol{K}}$, we have

$$\langle DV(u),\phi(u)\rangle_{\mathcal{Y}}=\langle D\Phi(u),\phi(u)\rangle_{\mathcal{Y}}+\langle\mathcal{C}_0^{-1/2}u,\mathcal{C}_0^{-1/2}\phi(u)\rangle_{\mathcal{Y}}=\mathrm{I}+\mathrm{II}, \tag{36}$$

where term II is understood as the white noise mapping. For term I, we have

$$\mathrm{I}=\langle\phi(\cdot),\boldsymbol{K}(u,\cdot)D\Phi(u)\rangle_{\mathcal{H}_{\boldsymbol{K}}}, \tag{37}$$

where the proposition (2) in Definition 4 is employed. For term II, we take $u_m\in\mathcal{H}^2$ such that $\lim_{m\to\infty}\|\mathcal{C}_0^{-\frac{1-s}{2}}(u_m-u)\|_{\mathcal{H}}=0$. It is clear to note that

$$\langle\mathcal{C}_0^{-1/2}u_m,\mathcal{C}_0^{-1/2}\phi(u)\rangle_{\mathcal{Y}}=\langle\mathcal{C}_0^{-1}u_m,\phi(u)\rangle_{\mathcal{Y}}=\langle\phi(\cdot),\boldsymbol{K}(u,\cdot)\mathcal{C}_0^{-1}u_m\rangle_{\mathcal{H}_{\boldsymbol{K}}}. \tag{38}$$

Because

$$\begin{aligned}
&|\langle\phi(\cdot),\boldsymbol{K}(u,\cdot)\mathcal{C}_0^{-1}u_m\rangle_{\mathcal{H}_{\boldsymbol{K}}}-\langle\phi(\cdot),\boldsymbol{K}(u,\cdot)\mathcal{C}_0^{-1}u\rangle_{\mathcal{H}_{\boldsymbol{K}}}|^2\\
&\quad\le\langle\phi(\cdot),\phi(\cdot)\rangle_{\mathcal{H}_{\boldsymbol{K}}}\langle\boldsymbol{K}(u,\cdot)\mathcal{C}_0^{-1}(u_m-u),\boldsymbol{K}(u,\cdot)\mathcal{C}_0^{-1}(u_m-u)\rangle_{\mathcal{H}_{\boldsymbol{K}}}\\
&\quad=\langle\phi(\cdot),\phi(\cdot)\rangle_{\mathcal{H}_{\boldsymbol{K}}}\langle\boldsymbol{K}(u,u)\mathcal{C}_0^{-1}(u_m-u),\mathcal{C}_0^{-1}(u_m-u)\rangle_{\mathcal{Y}}\\
&\quad\le\langle\phi(\cdot),\phi(\cdot)\rangle_{\mathcal{H}_{\boldsymbol{K}}}\langle\boldsymbol{K}(u,u)\mathcal{C}_0^{-1}(u_m-u),\mathcal{C}_0^{-1}(u_m-u)\rangle_{\mathcal{Y}}\\
&\quad\le C\langle\phi(\cdot),\phi(\cdot)\rangle_{\mathcal{H}_{\boldsymbol{K}}}\|\mathcal{C}_0^{-\frac{1-s}{2}}(u_m-u)\|_{\mathcal{H}}^2,
\end{aligned}$$

we find that $\lim_{m\to\infty}\langle\phi(\cdot),\boldsymbol{K}(u,\cdot)\mathcal{C}_0^{-1}u_m\rangle_{\mathcal{H}_{\boldsymbol{K}}}=\langle\phi(\cdot),\boldsymbol{K}(u,\cdot)\mathcal{C}_0^{-1}u\rangle_{\mathcal{H}_{\boldsymbol{K}}}$. Hence, letting $m\to\infty$ in (38), we have

$$\langle\mathcal{C}_0^{-1/2}u,\mathcal{C}_0^{-1/2}\phi(u)\rangle_{\mathcal{Y}}=\langle\phi(\cdot),\boldsymbol{K}(u,\cdot)\mathcal{C}_0^{-1}u\rangle_{\mathcal{H}_{\boldsymbol{K}}}. \tag{39}$$

Plugging (39) and (37) into (36), we obtain

$$\begin{aligned}\langle DV(u),\phi(u)\rangle_{\mathcal{Y}}&=\langle\phi(\cdot),\boldsymbol{K}(u,\cdot)D\Phi(u)+\boldsymbol{K}(u,\cdot)\mathcal{C}_0^{-1}u\rangle_{\mathcal{H}_{\boldsymbol{K}}}\\&=\langle\phi(\cdot),\boldsymbol{K}(u,\cdot)DV(u)\rangle_{\mathcal{H}_{\boldsymbol{K}}}.\end{aligned} \tag{40}$$

Next, let us calculate the second term on the right-hand side of (17). A simple calculation yields

$$\sum_{k=1}^{\infty} D_k \langle \phi(u), e_k\rangle_{\mathcal{Y}} = \sum_{k=1}^{\infty} D_k \langle \phi(\cdot), \boldsymbol{K}(u,\cdot)e_k\rangle_{\mathcal{H}_{\boldsymbol{K}}}. \tag{41}$$

Since

$$\begin{aligned} D_k \langle \phi(\cdot), \boldsymbol{K}(u,\cdot)e_k\rangle_{\mathcal{H}_{\boldsymbol{K}}} &= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\langle \phi(\cdot), \boldsymbol{K}(u+\epsilon\varphi_k,\cdot)e_k - \boldsymbol{K}(u,\cdot)e_k\rangle_{\mathcal{H}_{\boldsymbol{K}}} \\ &= \langle \phi(\cdot), D_k\boldsymbol{K}(u,\cdot)e_k\rangle_{\mathcal{H}_{\boldsymbol{K}}}, \end{aligned} \tag{42}$$

we have

$$\sum_{k=1}^{\infty} D_k \langle \phi(u), e_k\rangle_{\mathcal{Y}} = \left\langle \phi(\cdot), \sum_{k=1}^{\infty} D_k\boldsymbol{K}(u,\cdot)e_k \right\rangle_{\mathcal{H}_{\boldsymbol{K}}}. \tag{43}$$

Combining (40) and (43) with (17), we obtain

$$\mathcal{S}\phi(u) = \left\langle \phi(\cdot), -\boldsymbol{K}(u,\cdot)DV(u) + \sum_{k=1}^{\infty} D_k\boldsymbol{K}(u,\cdot)e_k \right\rangle_{\mathcal{H}_{\boldsymbol{K}}}. \tag{44}$$

Thus, the optimization problem (16) possesses a solution $\phi^*_{\boldsymbol{K}}(\cdot)$ satisfying

$$\phi^*_{\boldsymbol{K}}(\cdot) \propto \mathbb{E}_{u\sim\mu}\left[-\boldsymbol{K}(u,\cdot)DV(u) + \sum_{k=1}^{\infty} D_k\boldsymbol{K}(u,\cdot)e_k \right]. \tag{45}$$

Based on condition (33), we know that $D\phi^*_{\boldsymbol{K}}(u)$ belongs to $\mathcal{L}_1(\mathcal{X},\mathcal{Y})$ for each $u\in\mathcal{X}$, which completes the proof. □

Remark 11. The optimal $\phi^*_{\boldsymbol{K}}$ is given in (34) which is consistent with the finite-dimensional case. Since the first and second terms on the right-hand side of (34) are similar, we may just focus on the second term which is usually named as the repulsive force term. For each $u, v\in\mathcal{X}$, consider $\boldsymbol{K}(u,v) := K(u,v)\mathrm{Id}$ with $K(u,v) := \exp\left(-\frac{1}{h}\|u-v\|^2_{\mathcal{X}}\right)$. Then, we have

$$\begin{aligned} \sum_{k=1}^{\infty} D_{u_k}\boldsymbol{K}(u,v)e_k &= \sum_{k=1}^{\infty} \langle D_u K(u,v)e_k, \varphi_k\rangle_{\mathcal{X}} \\ &= \sum_{k=1}^{\infty} -\frac{2}{h}\langle u-v, \varphi_k\rangle_{\mathcal{X}} K(u,v)e_k. \end{aligned} \tag{46}$$

Projecting (46) onto one particular coordinate e_ℓ with $\ell\in\mathbb{N}$, we obtain

$$\begin{aligned} \left(\sum_{k=1}^{\infty} D_{u_k}\boldsymbol{K}(u,v)e_k\right)_{\ell} &= \left\langle \sum_{k=1}^{\infty} -\frac{2}{h}\langle u-v, \varphi_k\rangle_{\mathcal{X}} K(u,v)e_k, e_\ell \right\rangle_{\mathcal{Y}} \\ &= -\frac{2}{h}\langle u-v, \varphi_\ell\rangle_{\mathcal{X}} K(u,v), \end{aligned} \tag{47}$$

which is similar to the ℓth coordinate of $\nabla_{u^N} K(u^N, v^N)$ appearing in (9). Additionally, we mention that the assumption (33) given in Theorem 10 can be verified for many useful kernels. Detailed illustrations are provided in the supplementary material (supp.pdf [local/web 2.91MB]).

By Theorem 10, we can construct a series of transformations as follows,

$$T_\ell(u) = u + \epsilon_\ell \mathbb{E}_{u' \sim \mu_\ell}\left[-\boldsymbol{K}(u', u) DV(u') + \sum_{k=1}^{\infty} D_{(u')_k} \boldsymbol{K}(u', u) e_k \right] \tag{48}$$

with $\ell = 1, 2, \ldots$. In practice, we draw a set of particles $\{u_i^0\}_{i=1}^m$ from some initial measure, and then iteratively update the particles with an empirical version of the above transformation in which the expectation under μ_ℓ is approximated by the empirical mean of particles $\{u_i^\ell\}_{i=1}^m$ at the ℓth iteration. The iSVGD is summarized in Algorithm 2.

Algorithm 2. iSVGD.

```
Input: A target probability measure $\mu^{\boldsymbol{d}}$ that is absolutely continuous w.r.t the Gaussian measure $\mu_0 = \mathcal{N}(0, \mathcal{C}_0)$ with $\frac{d\mu^{\boldsymbol{d}}}{d\mu_0}(u) \propto \exp(-\Phi(u))$ and a set of particles $\{u_i^0\}_{i=1}^m$.
Output: A set of particles $\{u_i\}_{i=1}^m$ that approximates the target probability measure.
for iteration $\ell$ do
    $u_i^{\ell+1} \longleftarrow u_i^\ell + \epsilon_\ell \phi^*(u_i^\ell)$,
    where
    $\phi^*(u) = \frac{1}{m} \sum_{j=1}^{m} \left[ \boldsymbol{K}(u_j^\ell, u)(-D\Phi(u_j^\ell) - \mathcal{C}_0^{-1} u_j^\ell) + \sum_{k=1}^{\infty} D_{(u_j^\ell)_k} \boldsymbol{K}(u_j^\ell, u) e_k \right]$.
end for
```

3.3. iSVGD with precondition information.

In the supplementary material (supp.pdf [local/web 2.91MB]), the numerical experiments indicate that the SVGD without preconditioning operators converges slowly for some inverse problems of PDEs. By the finite-dimensional SVGD [58], it may accelerate the convergence and give reliable estimates efficiently by introducing preconditioning operators. For constructing the iSVGD with preconditioning operators, let us begin with a theorem concerning the change of variables.

THEOREM 12. *Let $\mathcal{X}$ and $\mathcal{Y}$ be two separable Hilbert spaces, and let $\mathcal{F}_0$ be an RKHS with a nonnegative $\mathcal{L}(\mathcal{Y})$-valued kernel $\boldsymbol{K}_0 : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$. Let $\tilde{\mathcal{X}}$ and $\tilde{\mathcal{Y}}$ be two separable Hilbert spaces, and $\mathcal{F}$ be the set of operators from $\tilde{\mathcal{X}}$ to $\tilde{\mathcal{Y}}$ given by*

$$\phi(u) = \boldsymbol{M}(u)\phi_0(t(u)) \quad \forall\ \phi_0 \in \mathcal{F}_0, \tag{49}$$

where $\boldsymbol{M} : \tilde{\mathcal{X}} \to \mathcal{L}(\mathcal{Y}, \tilde{\mathcal{Y}})$ is a fixed operator and is assumed to be an invertible operator for all $u \in \tilde{\mathcal{X}}$, and $t : \tilde{\mathcal{X}} \to \mathcal{X}$ is a fixed Fréchet differentiable one-to-one mapping. For all $\phi, \phi' \in \mathcal{F}$, we can identify a unique $\phi_0, \phi_0' \in \mathcal{F}_0$ such that $\phi(u) = \boldsymbol{M}(u)\phi_0(t(u))$ and $\phi'(u) = \boldsymbol{M}(u)\phi_0'(t(u))$. Define the inner product on $\mathcal{F}$ via $\langle \phi, \phi' \rangle_{\mathcal{F}} = \langle \phi_0, \phi_0' \rangle_{\mathcal{F}_0}$, and then $\mathcal{F}$ is also a vector-valued RKHS, whose operator-valued kernel is

$$\boldsymbol{K}(u, u') = \boldsymbol{M}(u')\boldsymbol{K}_0(t(u), t(u'))\boldsymbol{M}(u)^*, \tag{50}$$

where $\boldsymbol{M}(u)^*$ denotes the Hilbert space adjoint.

Proof. Let $\{(u_i, g_i)_{i=1,\ldots,N}\} \subset \tilde{\mathcal{X}} \times \tilde{\mathcal{Y}}$, and we have

$$\begin{aligned} \langle \boldsymbol{K}(u_i, u_j) g_i, g_j \rangle_{\tilde{\mathcal{Y}}} &= \langle \boldsymbol{M}(u_j)\boldsymbol{K}_0(t(u_i), t(u_j))\boldsymbol{M}(u_i)^* g_i, g_j \rangle_{\tilde{\mathcal{Y}}} \\ &= \langle \boldsymbol{K}_0(t(u_i), t(u_j))\boldsymbol{M}(u_i)^* g_i, \boldsymbol{M}(u_j)^* g_j \rangle_{\mathcal{Y}}. \end{aligned} \tag{51}$$

Then, the nonnegativity of $\boldsymbol{K}(\cdot,\cdot)$ follows from the nonnegative property of $\boldsymbol{K}_0(\cdot,\cdot)$. To prove the theorem, it suffices to verify the two conditions shown in Definition 4. For every $u, v \in \tilde{\mathcal{X}}$ and $g \in \tilde{\mathcal{Y}}$, we consider the operator $f(v) = \boldsymbol{K}(u,v)g = \boldsymbol{M}(v)\boldsymbol{K}_0(t(u),t(v))\boldsymbol{M}(u)^*g$. Because of $\boldsymbol{M}(u)^*g \in \mathcal{Y}$, we easily obtain

$$\boldsymbol{K}_0(t(u),t(v))\boldsymbol{M}(u)^*g \in \mathcal{F}_0.$$

According to (49), we conclude that $f(\cdot) \in \mathcal{F}$.

Next, let us verify the reproducing property of $\boldsymbol{K}(\cdot,\cdot)$. For every $u \in \tilde{\mathcal{X}}, g \in \tilde{\mathcal{Y}}$, and $\phi \in \mathcal{F}$, we have

$$\begin{aligned}\langle \phi(u), g\rangle_{\tilde{\mathcal{Y}}} &= \langle \boldsymbol{M}(u)\phi_0(t(u)), g\rangle_{\tilde{\mathcal{Y}}} = \langle \phi_0(t(u)), \boldsymbol{M}(u)^*g\rangle_{\mathcal{Y}} \\ &= \langle \phi_0(\cdot), \boldsymbol{K}_0(t(u),\cdot)\boldsymbol{M}(u)^*g\rangle_{\mathcal{F}_0} \\ &= \langle \boldsymbol{M}(\cdot)\phi_0(t(\cdot)), \boldsymbol{M}(\cdot)\boldsymbol{K}_0(t(u),t(\cdot))\boldsymbol{M}(u)^*g\rangle_{\mathcal{F}} \\ &= \langle \phi(\cdot), \boldsymbol{K}(u,\cdot)g\rangle_{\mathcal{F}},\end{aligned}$$

where the fourth equality follows from

$$\langle \phi, \phi'\rangle_{\mathcal{F}} = \langle \phi_0, \phi_0'\rangle_{\mathcal{F}_0}$$

with $\phi_0'(\cdot) = \boldsymbol{K}_0(t(u),\cdot)\boldsymbol{M}(u)^*g$. □

Now we present a key result, which characterizes the change of kernels when applying invertible transformations on the iSVGD trajectory.

Theorem 13. *Let $\mathcal{H}$, $\tilde{\mathcal{H}}$, $\mathcal{X}$, $\tilde{\mathcal{X}}$, $\mathcal{Y}$, and $\tilde{\mathcal{Y}}$ be separable Hilbert spaces satisfying $\mathcal{X} \subset \mathcal{Y}$, $\tilde{\mathcal{X}} \subset \tilde{\mathcal{Y}}$, $\mathcal{X} \subset \tilde{\mathcal{Y}}$, $\tilde{\mathcal{X}} \subset \mathcal{Y}$. Assume that Assumption 7 holds for the triples $(\mathcal{X},\mathcal{Y},\mathcal{H})$ and $(\tilde{\mathcal{X}},\tilde{\mathcal{Y}},\tilde{\mathcal{H}})$ with two fixed parameters $s \in [0,1]$, respectively. Let $T \in \mathcal{L}(\mathcal{Y},\tilde{\mathcal{Y}})$ and assume that T is a bounded operator when restricted to be an operator from $\mathcal{X}$ to $\tilde{\mathcal{X}}$. Let μ, $\mu^{\boldsymbol{d}}$ be two probability measures and $\tilde{\mu}$, $\tilde{\mu}^{\boldsymbol{d}}$ be the measures of $\tilde{u} = Tu$ when u is drawn from μ, $\mu^{\boldsymbol{d}}$, respectively. Introduce two Stein operators $\mathcal{S}$ and $\tilde{\mathcal{S}}$ as follows,*

$$\begin{aligned}\mathcal{S}\phi(u) &= \langle -DV(u), \phi(u)\rangle_{\mathcal{Y}} + \sum_{k=1}^{\infty} D_k\langle \phi(u), e_k\rangle_{\mathcal{Y}} \quad \forall\, u \in \mathcal{X}, \\ \tilde{\mathcal{S}}\tilde{\phi}(\tilde{u}) &= \langle -D_{\tilde{u}}V(T^{-1}\tilde{u}), \tilde{\phi}(\tilde{u})\rangle_{\tilde{\mathcal{Y}}} + \sum_{k=1}^{\infty} D_{(\tilde{u})_k}\langle \tilde{\phi}(\tilde{u}), \tilde{e}_k\rangle_{\tilde{\mathcal{Y}}} \quad \forall\, \tilde{u} \in \tilde{\mathcal{X}},\end{aligned}$$

where $\{e_k\}_{k=1}^{\infty}$ and $\{\tilde{e}_k\}_{k=1}^{\infty}$ are orthonormal bases in $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$, respectively. Then, we have

$$\mathbb{E}_{u\sim\mu}[\mathcal{S}\phi(u)] = \mathbb{E}_{u\sim\tilde{\mu}}[\tilde{\mathcal{S}}\tilde{\phi}(u)] \quad \textit{with } \phi(u) := T^{-1}\tilde{\phi}(Tu). \tag{52}$$

Therefore, in the asymptotics of infinitesimal step size ($\epsilon \to 0^+$), it is equivalent to running iSVGD with kernel $\boldsymbol{K}_0$ on $\tilde{\mu}$ and running iSVGD on μ with the kernel $\boldsymbol{K}(u,u') = T^{-1}\boldsymbol{K}_0(Tu,Tu')(T^{-1})^*$, in the sense that the trajectory of these two SVGD can be mapped to each other by the map T (and its inverse).

Proof. Let us introduce a mapping defined by $u' = f(u) = u + \epsilon\phi(u)$. Denote $f_{\#}\mu$ as the probability measure $\mu \circ f^{-1}$. Let $\tilde{u}' \sim T_{\#}(f_{\#}\tilde{\mu})$ which is obtained by

$$\begin{aligned}\tilde{u}' &= Tu' = T(u + \epsilon\phi(u)) = T(T^{-1}\tilde{u} + \epsilon\phi(T^{-1}\tilde{u})) \\ &= \tilde{u} + \epsilon T\phi(T^{-1}\tilde{u}) \\ &= \tilde{u} + \epsilon\tilde{\phi}(\tilde{u}),\end{aligned} \tag{53}$$

where we use the definition $\phi(u) = T^{-1}\tilde{\phi}(Tu)$ in (52). According to [39, Theorem 3.1] and [58, Theorem 3], we have $\mathbb{E}_{u^N \sim P^N_{\#}\mu}[\mathcal{S}^N\phi^N(u^N)] = \mathbb{E}_{u^N \sim P^N_{\#}\tilde{\mu}}[\tilde{\mathcal{S}}^N\tilde{\phi}^N(u^N)]$, where

$$\mathcal{S}^N\phi^N(u^N) = -\langle DV(u^N), \phi^N(u^N)\rangle_{\mathcal{Y}} + \sum_{k=1}^{N} D_k\langle \phi^N(u^N), e_k\rangle_{\mathcal{Y}},$$
$$\tilde{\mathcal{S}}^N\tilde{\phi}^N(\tilde{u}^N) = -\langle D_{\tilde{u}^N}V(T^{-1}\tilde{u}^N), \tilde{\phi}^N(\tilde{u}^N)\rangle_{\tilde{\mathcal{Y}}} + \sum_{k=1}^{N} D_{(\tilde{u}^N)_k}\langle \tilde{\phi}^N(\tilde{u}^N), \tilde{e}_k\rangle_{\tilde{\mathcal{Y}}}.$$

It is clear to note that there is no Jacobian matrix given by the transformation in $D_{\tilde{u}^N}V(T^{-1}\tilde{u}^N)$ since the Jacobian matrix does not depend on $\tilde{u}^N$ for linear mappings, i.e., the derivative is zero. Following the proof for Theorem 9, we take $N \to \infty$ and obtain $\mathbb{E}_{u\sim\mu}[\mathcal{S}\phi(u)] = \mathbb{E}_{u\sim\tilde{\mu}}[\tilde{\mathcal{S}}\tilde{\phi}(u)]$. From Theorem 12, when $\tilde{\phi}$ is in $\tilde{\mathcal{F}}$ with kernel $\boldsymbol{K}_0(u,u')$, ϕ is in $\mathcal{F}$ with kernel $\boldsymbol{K}(u,u')$. Therefore, maximizing $\mathbb{E}_{u\sim\mu}[\mathcal{S}\phi(u)]$ in $\mathcal{F}$ is equivalent to $\mathbb{E}_{u\sim\tilde{\mu}}[\tilde{\mathcal{S}}\tilde{\phi}(u)]$ in $\tilde{\mathcal{F}}$. This suggests that the trajectory of iSVGD on $\tilde{\mu}^{\boldsymbol{d}}$ with $\boldsymbol{K}_0$ and that on $\mu^{\boldsymbol{d}}$ with $\boldsymbol{K}$ are equivalent, which completes the proof. ∎

Remark 14. Similarly to the matrix-valued case [58], Theorem 13 suggests a conceptual procedure for constructing proper operator kernels to incorporate desirable preconditioning information. Differently from the finite-dimensional case, the map T is only allowed to be linear at this stage. For a nonlinear map, there is a Jacobian matrix in $\tilde{\mathcal{S}}^N\tilde{\phi}^N(\tilde{u}^N)$. It is difficult to analyze the limiting behavior of the Jacobian matrix related term. Practically, linear maps seem to be enough since even in the finite-dimensional case nonlinear maps will yield an unnatural algorithm [58].

In the last part of this subsection, we provide some examples of preconditioning operators that are frequently used in statistical inverse problems.

3.3.1. Fixed preconditioning operator. In section 5 of [16], the Langevin equation was considered by using $\mathcal{C}_0$ as a preconditioner, and an analysis was carried out for the pCN algorithm. For the Newton based iterative method, we usually take the inverse of the second-order derivative of the objective functional as the preconditioning operator [41]. Here, we consider a linear operator T that has similar properties to $\mathcal{C}_0^{-\frac{1-s}{2}}$. Specifically, we require

$$T \in \mathcal{L}(\mathcal{H}^{1-s}, \mathcal{H}) \cap \mathcal{L}(\mathcal{H}^{-1-s}, \mathcal{H}^{-2}). \tag{54}$$

Then, we specify the Hilbert space appearing in Theorem 12 as $\mathcal{X} = \mathcal{H}^{1-s}$, $\mathcal{Y} = \mathcal{H}^{-1-s}$, $\tilde{\mathcal{X}} = \mathcal{H}$, $\tilde{\mathcal{Y}} = \mathcal{H}^{-2}$ with $s \in [0,1]$. For the kernel $\boldsymbol{K}_0(\cdot,\cdot) : \tilde{\mathcal{X}} \times \tilde{\mathcal{X}} \to \tilde{\mathcal{Y}}$, we assume that

$$\sup_{\tilde{u}\in\mathcal{H}} \|\boldsymbol{K}_0(\tilde{u},\tilde{u})\|_{\mathcal{L}(\mathcal{H}^{-2})} < \infty. \tag{55}$$

It follows from Theorem 13 that we may use a kernel of the form

$$\boldsymbol{K}(u,u') := T^{-1}\boldsymbol{K}_0(Tu,Tu')(T^{-1})^*, \tag{56}$$

where $u, u' \in \mathcal{H}^{1-s}$. Obviously, the kernel $\boldsymbol{K}$ given above satisfies

$$\sup_{u\in\mathcal{H}^{1-s}} \|T^{-1}\boldsymbol{K}_0(Tu,Tu)(T^{-1})^*\|_{\mathcal{L}(\mathcal{H}^{-1-s})} < \infty. \tag{57}$$

As an example, we may take $\boldsymbol{K}_0$ to be the scalar-valued Gaussian RBF kernel composed with operator $\mathcal{C}_0^s$,

$$\boldsymbol{K}_0(u,u') := \exp\left(-\frac{1}{h}\|u-u'\|_{\mathcal{H}}^2\right)\mathcal{C}_0^s, \tag{58}$$

which yields

$$\boldsymbol{K}(u,u') = \exp\left(-\frac{1}{h}\|T(u-u')\|_{\mathcal{H}}^2\right)T^{-1}\mathcal{C}_0^s(T^{-1})^*, \tag{59}$$

where h is a bandwidth parameter. Define $\boldsymbol{K}_0^T(u,u') := \boldsymbol{K}_0(Tu,Tu')$. Let $\mathcal{P} := T^{-1}\mathcal{C}_0^s(T^{-1})^*$. By simple calculations, we find that the iSVGD update direction of the kernel in (56) is

$$\phi_{\boldsymbol{K}}^*(\cdot) = \mathcal{P}\mathbb{E}_{u\sim\mu}\left[\boldsymbol{K}_0^T(u,\cdot)(-D\Phi(u)-\mathcal{C}_0^{-1}u) + \sum_{k=1}^{\infty}D_k\boldsymbol{K}_0^T(u,\cdot)e_k\right] = \mathcal{P}\phi_{\boldsymbol{K}_0^T}^*, \tag{60}$$

which is a linear transform of the iSVGD update direction of the kernel $\boldsymbol{K}_0^T$ with the operator $T^{-1}\mathcal{C}_0^s(T^{-1})^*$.

3.3.2. The $\mathcal{C}_0$ operator. Choosing $T := \mathcal{C}_0^{-\frac{1-s}{2}}$, we can see that condition (54) holds. Given the Kernel $\boldsymbol{K}_0$ in (58), the kernel $\boldsymbol{K}$ defined in (59) can be written as

$$\boldsymbol{K}(u,u') = \exp\left(-\frac{1}{h}\|\mathcal{C}_0^{-\frac{1-s}{2}}(u-u')\|_{\mathcal{H}}^2\right)\mathcal{C}_0.$$

The operator $\mathcal{P}$ used in (60) is just $\mathcal{C}_0$. If there is only one particle, the iSVGD update direction is then reduced to $\phi_{\boldsymbol{K}}^*(\cdot) = \mathcal{C}_0(D\Phi(u) + \mathcal{C}_0^{-1}u)$.

3.3.3. The Hessian operator. For statistical inverse problems, the forward operator $\mathcal{G}$ is usually nonlinear, e.g., the inverse medium scattering problem [26, 27]. Around each particle u_i with $i = 1, 2, \ldots, m$, the forward map can be approximated by the linearized map

$$\mathcal{G}(u) \approx \mathcal{G}(u_i) + D\mathcal{G}(u_i)(u-u_i). \tag{61}$$

Assume that the potential function Φ takes the form $\Phi(u) = \frac{1}{2}\|\Sigma^{-1/2}(\mathcal{G}(u)-d)\|_{\ell^2}^2$, where Σ is a positive-definite matrix. Using the approximate formula (61), we have

$$V(u) \approx \tilde{V}(u) := \frac{1}{2}\|\Sigma^{-1/2}(D\mathcal{G}(u_i)u - D\mathcal{G}(u_i)u_i + \mathcal{G}(u_i) - d)\|_{\ell^2}^2 + \frac{1}{2}\|\mathcal{C}_0^{-1/2}u\|_{\mathcal{H}}^2.$$

It follows from a simple calculation that $D^2\tilde{V}(u_i) = D\mathcal{G}(u_i)^*\Sigma^{-1}D\mathcal{G}(u_i) + \mathcal{C}_0^{-1}$. For the Newton-type iterative method, we can take the linear transformation $T = \mathcal{C}_0^{s/2}(\frac{1}{m}\sum_{i=1}^m(D\mathcal{G}(u_i)^*\Sigma^{-1}D\mathcal{G}(u_i) + \mathcal{C}_0^{-1}))^{1/2}$. If $\mathcal{G}$ is a linear operator (e.g., the examples in [25]), it is easy to verify condition (54). For nonlinear problems, it is necessary to employ the regularity properties of the direct problems, which is beyond the scope of this work. Hence we will not verify this condition in this paper and leave it as a future work. With this choice of T, the kernel (59) and the iSVGD update direction (60) can be easily obtained. If there is only one particle, the iSVGD update direction is degenerated to the usual Newton update direction when evaluating the MAP estimate.

3.3.4. Mixture preconditioning. Using a fixed preconditioning operator, we cannot specify different preconditioning operators for different particles. Inspired by the mixture precondition [58], we propose an approach to achieve pointwise preconditioning. The idea is to use a weighted combination of several linear preconditioning operators. This involves leveraging a set of anchor points $\{v_\ell\}_{\ell=1}^m$, each of which is associated with a preconditioning operator T_ℓ (e.g.,

$$T_\ell = \mathcal{C}_0^{s/2}(D\mathcal{G}(v_\ell)^*\Sigma^{-1}D\mathcal{G}(v_\ell) + \mathcal{C}_0^{-1})^{1/2}).$$

In practice, the anchor points $\{v_\ell\}_{\ell=1}^m$ can be set to be the same as the particles $\{u_i\}_{i=1}^m$. We then construct a kernel by $\boldsymbol{K}(u,u') = \sum_{\ell=1}^m \boldsymbol{K}_\ell(u,u')w_\ell(u)w_\ell(u')$, where

$$\boldsymbol{K}_\ell(u,u') := T_\ell^{-1}\boldsymbol{K}_0(T_\ell u, T_\ell u')(T_\ell^{-1})^*, \tag{62}$$

and $w_\ell(u)$ is a positive scalar-valued function that determines the contribution of kernel $\boldsymbol{K}_\ell$ at point u. Here $w_\ell(u)$ should be viewed as a mixture probability, and hence should satisfy $\sum_{\ell=1}^m w_\ell(u) = 1$ for all u. In our empirical studies, we take

$$w_\ell(u) = \frac{\exp\Big(-\frac{1}{2}\|T_\ell(u-v_\ell)\|_{\mathcal{H}}^2\Big)}{\sum_{\ell'=1}^m \exp\Big(-\frac{1}{2}\|T_{\ell'}(u-v_{\ell'})\|_{\mathcal{H}}^2\Big)}. \tag{63}$$

In this way, each point u is mostly influenced by the anchor point closest to it, which allows us to apply different preconditioning for different points. In addition, the iSVGD update direction has the form

$$\begin{aligned}\phi_{\boldsymbol{K}}^*(\cdot) = \sum_{\ell=1}^m w_\ell(\cdot)\mathbb{E}_{u\sim\mu}\Bigg[&-w_\ell(u)\boldsymbol{K}_\ell(u,\cdot)(D\Phi(u) + \mathcal{C}_0^{-1}u)\\ &+\sum_{k=1}^\infty D_k(w_\ell(u)\boldsymbol{K}_\ell(u,\cdot)e_k)\Bigg],\end{aligned} \tag{64}$$

which is a weighted sum of a number of iSVGD update directions with linear preconditioning operators. The implementation details of (64) are given in the supplementary material (supp.pdf [local/web 2.91MB]).

Remark 15. For the kernel defined above, the particles should belong to the Hilbert space $\mathcal{H}^{1-s}$. Based on the studies the finite-dimensional problems [58], we may let the parameter s be equal to 0. However, when the parameter $s = 0$, each particle u_i belongs to $\mathcal{H}^1$ which is the Cameron–Martin space of the prior measure. By the classical Gaussian measure theory [15], we know that $\mathcal{H}^1$ has zero measure. This fact implies that all of the particles belong to a set with zero measure, which may lead to too concentrated particles and deviates from our purpose. Hence we should choose $s > 0$ to ensure the effectiveness of the SVGD sampling algorithm. These observations are illustrated by our numerical experiments in section 4.

3.4. Some insights about iSVGD. We have constructed the well-defined iSVGD algorithms with or without preconditioning operators, which is the first step to extend the finite-dimensional SVGD to the infinite-dimensional space. Some mathematical studies have been carried out for the finite-dimensional SVGD, e.g., gradient flow on probability space [38] and mean field limit theory related to the macroscopic behavior [42]. These results provide in-depth understandings of the SVGD algorithm and motivate many new algorithms [37]. In this subsection, we intend to provide a preliminary mathematical study on the iSVGD under a simpler setting.

We consider the kernel operator $\boldsymbol{K}(u,v) := K(\|u-v\|_{\mathcal{H}})\mathrm{Id}$ with $u, v \in \mathcal{H}$ and $K(\cdot)$ being a scalar function. Let m be the sample number and $V(u)$ be defined in (15). Similarly to the finite-dimensional case, the iterative procedure in Algorithm 2 can be viewed as a particle system

$$\begin{aligned} &\frac{d}{dt}u_i(t) = -(\tilde{D}\boldsymbol{K} * \mu_m(t))(u_i(t)) - (\boldsymbol{K} * DV\mu_m(t))(u_i(t)), \\ &\mu_m(t) = \frac{1}{m}\sum_{j=1}^{m}\delta_{u_j(t)}, \\ &u_i(0) = u_i^0, \quad i = 1, 2, \ldots, m, \end{aligned} \tag{65}$$

where $\{u_i^0\}_{i=1}^m$ are the initial particles, $\delta_{u_i(t)}$ denotes the Dirac measure concentrated on $u_i(t)$ with $i = 1, 2, \ldots, m$, "$*$" denotes the usual convolution operator, and $\tilde{D}\boldsymbol{K}(u-v) = \sum_{k=1}^{\infty} D_{u_k}\boldsymbol{K}(u-v)e_k$. For convenience, we write the two convolution terms in the following forms:

$$\begin{aligned} (\tilde{D}\boldsymbol{K} * \mu_m(t))(u_i(t)) &= \frac{1}{m}\sum_{j=1}^{m}\tilde{D}\boldsymbol{K}(u_i(t) - u_j(t)), \\ (\boldsymbol{K} * DV\mu_m(t))(u_i(t)) &= \frac{1}{m}\sum_{j=1}^{m}\boldsymbol{K}(u_i(t) - u_j(t))DV(u_j(t)). \end{aligned}$$

Similarly, we consider the weak form equation about the measure-valued function,

$$\begin{aligned} &\frac{d}{dt}\langle \mu(t), \varphi\rangle = \langle \mu(t), L(\mu(t))\varphi\rangle, \\ &\mu(0) = \nu, \end{aligned} \tag{66}$$

where ν is the probability measure employed to generate initial particles, φ is the test function, and

$$L(\mu(t))\varphi = \langle \tilde{D}\boldsymbol{K} * \mu(t), D\varphi\rangle_{\mathcal{H}} + \langle \boldsymbol{K} * DV\mu(t), D\varphi\rangle_{\mathcal{H}}. \tag{67}$$

Let $W^{1,2}(\mathcal{H}, \mu)$ be the usual Sobolev space defined for a Gaussian measure μ [47].

THEOREM 16. *Let μ_0 and Φ be the prior measure and potential function defined in* (1)*, respectively. Assume $K(\cdot) \in W^{1,2}(\mathcal{H}, \mu_0)$ and $e^{-\Phi(\cdot;\boldsymbol{d})} \in L^2(\mathcal{H}, \mu_0)$. Then, the posterior measure $\mu^{\boldsymbol{d}}$ defined in* (1) *is an invariant solution to* (66)*, i.e., when $\nu := \mu^{\boldsymbol{d}}$, the solution $\mu(t)$ of* (66) *is equal to $\mu^{\boldsymbol{d}}$.*

The proof is given in the supplementary material (supp.pdf [local/web 2.91MB]). Clearly, this theorem holds in the finite-dimensional setting. We point out that the integration by parts may not hold for the infinite-dimensional case. In the finite-dimensional setting, the analysis of the corresponding particle system (65) and (66) have been given recently in [42]. It is sophisticated to define meaningful solutions for the above interacting particle system (65) and the measure-valued function equation (66), which are beyond the scope of this study and are left for future work. One of the major difficulties for the infinite-dimensional case is that $\mathcal{C}_0^{-1}$ (the precision operator of the prior measure) is usually an unbounded operator [16]. Nearly all of the estimates presented in [42] for the finite-dimensional case cannot be adopted for the infinite-dimensional setting.

Numerical experiments indicate that the SVGD without preconditioning operators can hardly provide accurate estimates for some inverse problems. The SVGD with preconditioning operators can accelerate the convergence and give reliable estimates efficiently. In addition, the unboundedness issue induced by the precision operator $\mathcal{C}_0^{-1}$ may be overcome by introducing preconditioning operators. A detailed analysis of the iSVGD with preconditioning operators may be a good starting point for future theoretical studies.

At the end of this subsection, we mention a critical difference between finite- and infinite-dimensional theories. It follows from Theorem 2.7 in [42] and Theorem 1.1 in [57] that the empirical measure constructed by particles in finite-dimensional SVGD can approximate the continuous counterpart with accuracy ϵ when the number of particles are of order $O(\epsilon^d)$, where d is the discrete dimension. Obviously, an infinite number of particles is needed if the dimension d goes to infinity, which indicates that the infinite-dimensional theory may be meaningless.

The above statement explains that not every finite-dimensional setting can be meaningfully generalized to the infinite-dimensional space. The assumption on prior measure is important for the infinite-dimensional theory (the current assumption may be slightly relaxed, e.g., the Besov-type measure). According to the general analysis for the convergence and concentration of empirical measures given in [34], we believe that the prior measures used here can be approximated by the empirical measures under the Wasserstein distance on infinite-dimensional Hilbert space. Specifically, the estimate of the convergence speed is not relevent to the dimension when considering some finite-dimensional spaces as the projected infinite-dimensional space. If a theorem similar to Theorem 2.7 in [42] for the system (65)–(66) can be proved, we are able to confirm that the particles obtained by iSVGD can approximate the posterior measure for certain accuracy with particle numbers independent of the discrete dimension. However, it is higly nontrivial to carry out an in-depth study of the system (65)–(66) and is beyond the scope of the current work. In subsection SM6.3 of the supplementary material (supp.pdf [local/web 2.91MB]), we give a numerical illustration to address this issue.

4. Applications.

The proposed framework is valid for Bayesian inverse problems governed by any systems of PDEs. Due to the page limitation, we present one example of an inverse problem governed by the steady state Darcy flow equation. The second example concerns an inverse problem of the Helmholtz equation and is given in the supplementary material (supp.pdf [local/web 2.91MB]).

Consider the following PDE model,

$$\begin{aligned} -\nabla\cdot(e^u\nabla w) &= f \quad \text{in } \Omega, \\ w &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{68}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded Lipschitz domain, $f(x)$ denotes the sources, and $e^{u(x)}$ describes the permeability of the porous medium. This model is used as a benchmark problem in many works, e.g., the pCN algorithm [13] and the sequential Monte Carlo method [2]. We will compare the performance of the proposed iSVGD approach with the pCN [13] and the randomized MAP (rMAP) methods [59].

4.1. Basic settings and finite-element discretization.

For numerical implementations, it is essential to compute all of the related gradients and Hessian operators before discretization (i.e., pushing the discretization to the last step). A direct calculation yields the gradient and Hessian operators of the operator-valued kernel, but the adjoint method [41] needs to be employed for the potential Φ involving PDEs. More discussions on finite- and infinite-dimensional approaches can be found in the supplementary material (supp.pdf [local/web 2.91MB]), which might be helpful for readers who are not familiar with the infinite-dimensional approach. Let $\mathcal{F}$ be the solution operator that maps the parameter u to the solution of (68), and $\mathcal{M}$ be the measurement operator defined as $\boldsymbol{d} = \mathcal{M}(w) = (\ell_{x_1}(w), \ell_{x_2}(w), \dots, \ell_{x_{N_d}}(w))^T$, where

$$\ell_{x_j}(w) = \int_\Omega \frac{1}{2\pi\delta^2} e^{-\frac{1}{2\delta^2}\|x-x_j\|^2} w(x)dx \tag{69}$$

with $\delta > 0$ being a sufficiently small number and $x_i \in \Omega$ for $i = 1, \dots, N_d$. The forward map can be defined as $\mathcal{G} := \mathcal{M} \circ \mathcal{F}$, and the problem can be written in the abstract form $\boldsymbol{d} = \mathcal{G}(u) + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathrm{Id})$. Then we have $\Phi(u) = \frac{1}{2\sigma^2}\|\mathcal{M}(w) - \boldsymbol{d}\|^2$. The gradient $D\Phi(u)$ acting in any direction $\tilde{u}$ is given by

$$\langle D\Phi(u), \tilde{u}\rangle = \int_\Omega \tilde{u} e^u \nabla w \cdot \nabla p dx, \tag{70}$$

where the adjoint state p satisfies the *adjoint equation*

$$\begin{aligned} -\nabla \cdot (e^u \nabla p) &= -\frac{1}{\sigma^2}\sum_{j=1}^{N_d} \frac{1}{2\pi\delta^2} e^{-\frac{1}{2\delta^2}\|x-x_j\|^2} (\ell_{x_j}(w) - d_j) \quad \text{in } \Omega, \\ p &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{71}$$

The Hessian acting in direction $\tilde{u}$ and $\hat{u}$ reads

$$\begin{aligned} \langle\langle D^2\Phi(u), \hat{u}\rangle, \tilde{u}\rangle = &\int_\Omega \hat{u}\tilde{u} e^u \nabla w \cdot \nabla p dx + \int_\Omega \tilde{u} e^u \nabla w \cdot \nabla \hat{p} dx \\ &+ \int_\Omega \tilde{u} e^u \nabla p \cdot \nabla \hat{w} dx, \end{aligned} \tag{72}$$

where the state $\hat{w}$ satisfies the *incremental forward equation*

$$\begin{aligned} -\nabla \cdot (e^u \nabla \hat{w}) &= \nabla \cdot (\hat{u} e^u \nabla w) \quad \text{in } \Omega, \\ \hat{w} &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{73}$$

and the state $\hat{p}$ satisfies *the incremental adjoint equation*

$$\begin{aligned} -\nabla \cdot (e^u \nabla \hat{p}) &= \nabla \cdot (\hat{u} e^u \nabla p) - \frac{1}{2\pi\delta^2\sigma^2}\sum_{j=1}^{N_d} \hat{w} e^{-\frac{1}{2\delta^2}\|x-x_j\|^2} \quad \text{in } \Omega, \\ \hat{p} &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{74}$$

In experiments, we choose Ω to be a rectangular domain $\Omega = [0,1]^2 \subset \mathbb{R}^2$, set $\mathcal{H} = L^2(\Omega)$, and consider the prior measure $\mu_0 = \mathcal{N}(u_0, \mathcal{C}_0)$ with the mean function u_0 and the covariance operator $\mathcal{C}_0 := A^{-2}$, where $A = \alpha(I - \Delta)$ ($\alpha > 0$) with the domain of A given by $D(A) := \{u \in H^2(\Omega) : \frac{\partial u}{\partial \boldsymbol{n}} = 0 \text{ on } \partial\Omega\}$. Here, $H^2(\Omega)$ is the usual Sobolev space. Assume that the mean function u_0 resides in the Cameron–Martin space of μ_0.

Based on (70) and (72), we can prove the following results, which satisfy Assumptions 5. The proof is given in the supplementary material (supp.pdf [local/web 2.91MB]).

THEOREM 17. *Let $H^{-1}(\Omega)$ be the usual Sobolev space with the regularity index -1. Assume $\mathcal{X} = \mathcal{H}^{1-s}$ with the parameter $s < 0.5$, and then we have*

$$\begin{aligned} 0 \leq \Phi(u) &\leq C(1+\|f\|_{H^{-1}})^2 e^{2\|u\|_{\mathcal{X}}}, \\ \|D\Phi(u)\|_{\mathcal{X}^*} &\leq C(1+\|f\|_{H^{-1}})^2 e^{4\|u\|_{\mathcal{X}}}, \\ \|D^2\Phi(u)\|_{\mathcal{L}(\mathcal{X},\mathcal{X}^*)} &\leq C(1+\|f\|_{H^{-1}})^2 e^{6\|u\|_{\mathcal{X}}}. \end{aligned}$$

In the following, we use the Gaussian kernel, i.e., $\boldsymbol{K}(u,u') = \exp(-\frac{1}{h}\|u-u'\|_{\mathcal{H}}^2)$ for the iSVGD without preconditioning operators. For numerical examples with preconditioning operators, we employed the kernel given in subsection 3.3.4.

For finite-dimensional approximations, we consider a finite-dimensional subspace V_h of $L^2(\Omega)$ originating from the finite-element discretization with the continuous Lagrange basis functions $\{\phi_j\}_{j=1}^n$, which correspond to the nodal points $\{x_j\}_{j=1}^n$, such that $\phi_j(x_i) = \delta_{ij}$ for $i, j \in \{1, \ldots, n\}$. Instead of statistically inferring parameter functions $u \in L^2(\Omega)$, we consider the approximation $u_h = \sum_{j=1}^n u_j\phi_j \in V_h$. Under this finite-dimensional approximation, we can employ the numerical method provided in [4] to discretize the prior, and construct finite-dimensional approximations of the Gaussian approximation of the posterior measure. Based on our analysis in subsection 3.3, we need to calculate the fractional powers of the operator $\mathcal{C}_0$. Here, we employ the matrix transfer technique (MTT) [6]. The main idea of the MTT is to indirectly discretize a fractional Laplacian using a discretization of the standard Laplacian. As discussed in [4], the operator M is taken as

$$M = (M_{ij})_{i,j=1}^n \quad \text{and} \quad M_{ij} = \int_\Omega \phi_i(x)\phi_j(x)dx, \quad i,j \in \{1,\ldots,n\}. \tag{75}$$

The matrix $M^{1/2}$ is approximated by the diagonal matrix $\text{diag}(M_{11}^{1/2}, \ldots, M_{nn}^{1/2})$.

Finally, we mention that the finite element discretization is implemented by employing the open software FEniCS (Version 2019.1.0) [40]. All programs were run on a personal computer with Intel(R) Core(TM) i7-7700 at 3.60 GHz (CPU), 32 GB (memory), and Ubuntu 18.04.2 LTS (OS).

4.2. Numerical results. In the experiments, the noise level is fixed to be 1% since the goal is to test algorithms rather than demonstrate the Bayesian modeling. We compare the iSVGD with the mixture preconditioning operator (iSVGDMPO) with the pCN sampling algorithm [16] and the rMAP algorithm [59]. Since the rMAP sampling algorithm is not accurate for nonlinear problems, we choose $\alpha = 0.5$ in the prior probability measure. It should be mentioned that we choose the anchor points in the iSVGDMPO just to be the same as the particles and the anchor points will be updated during the iterations. The initial particles of the iSVGD are generated from a probability measure by using the method proposed in [4].

For the current settings, the gradient descent based method hardly seems able to find appropriate solutions in reasonable iterative steps. Hence, the optimization method with preconditioning operators, e.g., the Newton-conjugate gradient method, is employed. The term $\mathbb{E}_{u'\sim\mu_\ell}[\boldsymbol{K}(u',u)DV(u')]$ in (48) is an averaged gradient descent component in the whole iterative term, which drives all of the particles to be concentrated. We anticipate that Algorithm 2 cannot work well due to the inefficiency of the gradient descent algorithm. Due to the page limitation, numerical results are given in the supplementary material (supp.pdf [local/web 2.91MB]), which show that Algorithm 2 does not perform well in some cases. This is one of the main motivations for us to study the iSVGD with preconditioning operators.

We compare the iSVGD with the iSVGDMPO with those obtained by the pCN and rMAP sampling algorithms. As illustrated in Remark 15, the parameter s should not be zero. Intuitively, the particles should belong to a space with probability approximately equal to one under the prior measure μ_0. By the Gaussian measure theory [15], we may take $s > 0.5$ since $\mu_0(\mathcal{H}^{1-s}) = 1$ for any $s > 0.5$. Since the posterior measure is usually concentrated on a small support set of the prior measure, the parameter s should be slightly smaller than 0.5. Thus, we set $s = 0.3$ or 0.4 in our examples. Usually, the initial particles are scattered, and the variances of the initial particles are larger than the final particles obtained by the iSVGDMPO. We design the following adaptive empirical strategy for s,

$$s = -0.5\frac{\|\mathrm{var}\|_{\ell^2}}{\|\mathrm{var}_0\|_{\ell^2}} + 0.5, \tag{76}$$

where var is the current estimated variance, var_0 is the estimated variance of the initial particles, and $\|\cdot\|_{\ell^2}$ is the usual ℓ^2-norm. Obviously, for the initial particles, we have $s = 0$. The particles are forced to be concentrated. When the variance is reduced, the parameter s approaches 0.5 to avoid that the particles are concentrated on a set with zero measure. Since the pCN is a dimension independent MCMC-type sampling algorithm, we take the results obtained by the pCN as the baseline (accurate estimate). To make sure that the pCN algorithm yields an accurate estimate, we iterate 10^6 steps and withdraw the first 10^5 samples. Several different step sizes are tried and the traces of some parameters are plotted, and then the most reliable one is picked as the baseline.

In Figure 1, we show the estimated variances obtained by the iSVGDMPO (blue solid line), rMAP (green dotted line), and the pCN (orange dashed line) sampling algorithms. The estimated variances of the iSVGDMPO are shown for $s = 0$ and $s = 0.4$ on the left and in the middle, respectively. On the right, we exhibit the estimated variances when the empirical adaptive strategy (76) is employed. As expected, the estimated variances are too small when $s = 0$, which indicates that the particles are concentrated on a small set. Choosing $s = 0.4$ or using the empirical strategy, we obtain similar estimates, which is more similar to the baseline obtained by the pCN compared with the estimates obtained by the rMAP.

One important question arises: How does s influence the convergence of the iSVGDMPO? The detailed numerical comparisons are given in the supplementary material (supp.pdf [local/web 2.91MB]). Here we state the conclusions: The convergence speeds are similar for $s = 0.4$ and the adaptively chosen s. When specifying $s = 0.5$, the variances will gradually approach the background truth, but the con-

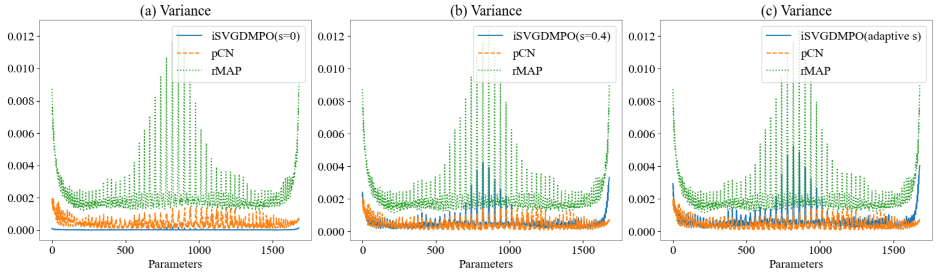

FIG. 1. *The comparison of the variances estimated by the pCN, rMAP, iSVGDMPO with different s.* (a)*: $s = 0$;* (b)*: $s = 0.4$;* (c)*: adpatively chosen s.*

vergence speed seems much slower than $s = 0.4$ or the adaptively chosen s. In the following numerical experiments, we use the empirical adaptive strategy to specify the parameter s.

In addition, we provide videos to exhibit the dynamic changing procedure of the estimated variances in the supplementary material (see compare_s_0_DarcyFlow.mp4 [local/web 21.5MB], compare_s_0.4_DarcyFlow.mp4 [local/web 20.9MB], compare_s_alter_DarcyFlow.mp4 [local/web 21.6MB], compare_s_0_Helmholtz.mp4 [local/web 9.67MB], compare_s_0.4_Helmholtz.mp4 [local/web 9.38MB], and compare_s_alter_Helmholtz.mp4 [local/web 9.25MB]). The update perturbation with and without the repulsive force term are exhibited. These videos can further illustrate our theoretical findings. We can see that the repulsive force terms indeed prevent the particles from being over-concentrated.

Apart from the parameter s, how many samples should be taken to guarantee a stable statistical quantity estimate is important for using the iSVGDMPO. When the particle number is too small, we cannot obtain reliable estimates. However, the computational complexity increases when the particle number increases. In Figure 2, we show the estimated variances when the particle number equals 10, 20, 30, 40, and 50. Denote by m the number of samples. On the left in Figure 2, we show the results obtained when $m = 10, 20, 30$. Obviously, when $m = 10$, the estimated variances are significantly smaller than those obtained when $m = 20, 30$. On the right in Figure 2, we find that the estimated variances are similar when $m = 30, 40, 50$. Hence, it is enough for our numerical examples to take $m = 20$ or 30, which attains a balance between efficiency and accuracy. So far, we have only compared the variances with different parameters in the iSVGDMPO. In the following, qualitative and quantitative comparisons of other statistical quantities are provided to illustrate the effectiveness of the iSVGDMPO.

Now, we specify the sampling number $m = 30$ and set the parameter s by the proposed empirical strategy (76). In Figure 3, we show the background truth and the estimated mean and variance functions obtained by the pCN, rMAP, and iSVGDMPO, respectively. The iterative number of the iSVGDMPO is set to be 30. From the first line, we observe that the mean functions obtained by the rMAP and iSVGDMPO are similar, which are slightly smoother than the one obtained by the pCN algorithm. This may be caused by the inexact matrix-free Newton-conjugate gradient algorithm [4]. As investigated in [59], many more powerful Newton-type algorithms can be employed to improve the performance both of the rMAP and iSVGDMPO. For the variances, the iSVGDMPO gives more reliable estimates compared with the rMAP,

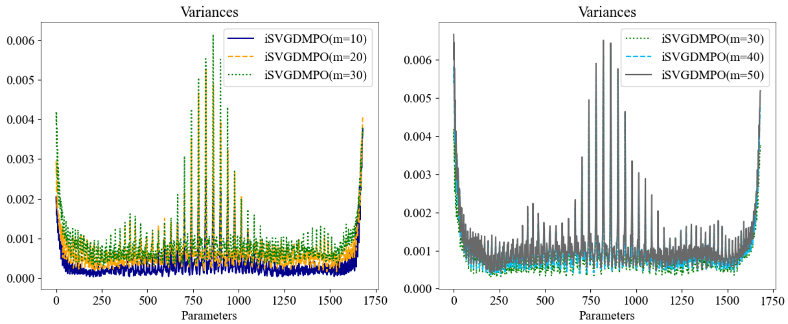

FIG. 2. *The comparison of the variances estimated by the iSVGDMPO with* $s = 10, 20, 30, 40, 50$.

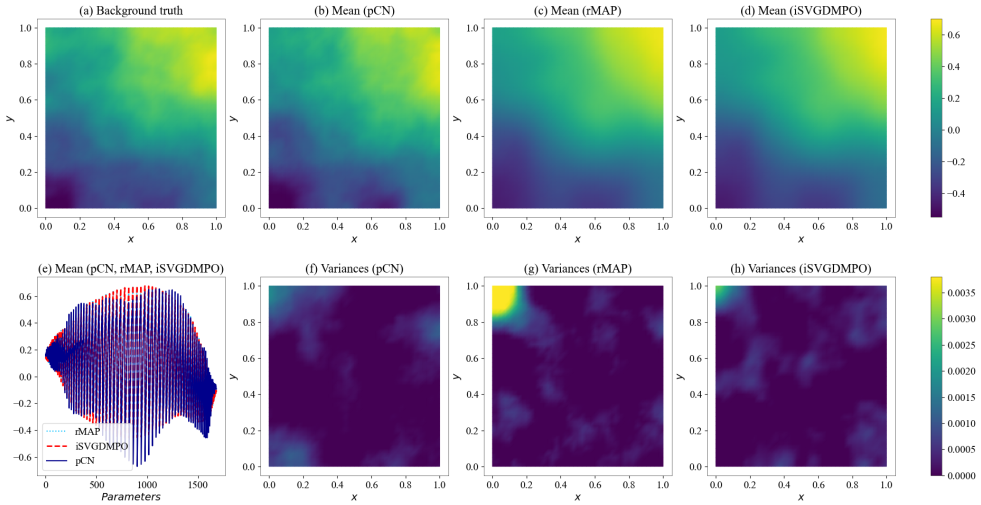

FIG. 3. *The background truth and the estimated mean and variance functions by the pCN, rMAP, and iSVGDMPO.* (a)*: The background truth;* (b)*: The estimated mean function by the pCN;* (c)*: The estimated mean function by the rMAP;* (d)*: The estimated mean function by the iSVGDMPO;* (e)*: The estimated mean function on mesh points by the pCN (blue solid line), rMAP (light blue dotted line), and iSVGDMPO (red dashed line);* (f)*: The estimated variances by the pCN;* (g)*: The estimated variances by the rMAP;* (h)*: The estimated variances by the iSVGDMPO.*

as can be seen from Figures 3(f), (g), and (h).

Next, we provide some more comparisons of statistical quantities between the results obtained by the pCN, rMAP, and iSVGDMPO. The samples are discretization of functions. As introduced in [49], the mean, variance, and covariance functions are the main statistics for functional data. The variance function denoted by $\text{var}_u(x)$ can be defined as $\text{var}_u(x) = \frac{1}{m}\sum_{i=1}^{m}(u_i(x) - \bar{u}(x))^2$, where $x \in \Omega$ is a point residing in the domain Ω, $\bar{u}$ is the mean function, and m is the sample number. The covariance function can be defined as $\text{cov}_u(x_1, x_2) = \frac{1}{m-1}\sum_{i=1}^{m}(u_i(x_1) - \bar{u}(x_1))(u_i(x_2) - \bar{u}(x_2))$, where $x_1, x_2 \in \Omega$ and $m, \bar{u}$ are defined as in $\text{var}_u(x)$. For simplicity, we compute these quantities on the mesh points and exhibit the results in Figure 4. In all of the subfigures in Figure 4, the estimates obtained by the pCN, rMAP, and iSVGDMPO are drawn in blue solid line, gray dotted line, and red dashed line, respectively. In Figure 4(a), we show the variance function calculated on all of the mesh points, i.e., $\{\text{var}_u(x_i)\}_{i=1}^{N_g}$ (N_g is the number of mesh points). In Figures 4(c) and (e), we show the covariance function calculated on the pairs of points $\{(x_i, x_{i+50})\}_{i=1}^{N_g-50}$ and $\{(x_i, x_{i+100})\}_{i=1}^{N_g-100}$, respectively. Compared with the estimates given by the rMAP, we can find that the estimates obtained by the iSVGDMPO are visually more similar to the estimates provided by the pCN. In Figures 4(b), (d), and (f), we provide the same estimates shown in (a), (c), and (e) with points indexing from 1000 to 1200, which give more detailed comparisons. The results also confirm that the iSVGDMPO provides more similar estimates to the pCN.

In addition, a quantitative comparison among the pCN, rMAP, and iSVGDMPO is given in Table 1. We compute the ℓ^2-norm differences of the variance and covariance functions on the mesh points obtained by the pCN, rMAP, and iSVGDMPO. In the table, the notation $\text{cov}_u(x_i, x_{i+k})$ ($k = 10, 20, \ldots, 110$) means the covariance function values on the pair of mesh points $\{(x_i, x_{i+k})\}_{i=1}^{N_g}$. The numbers below this notation

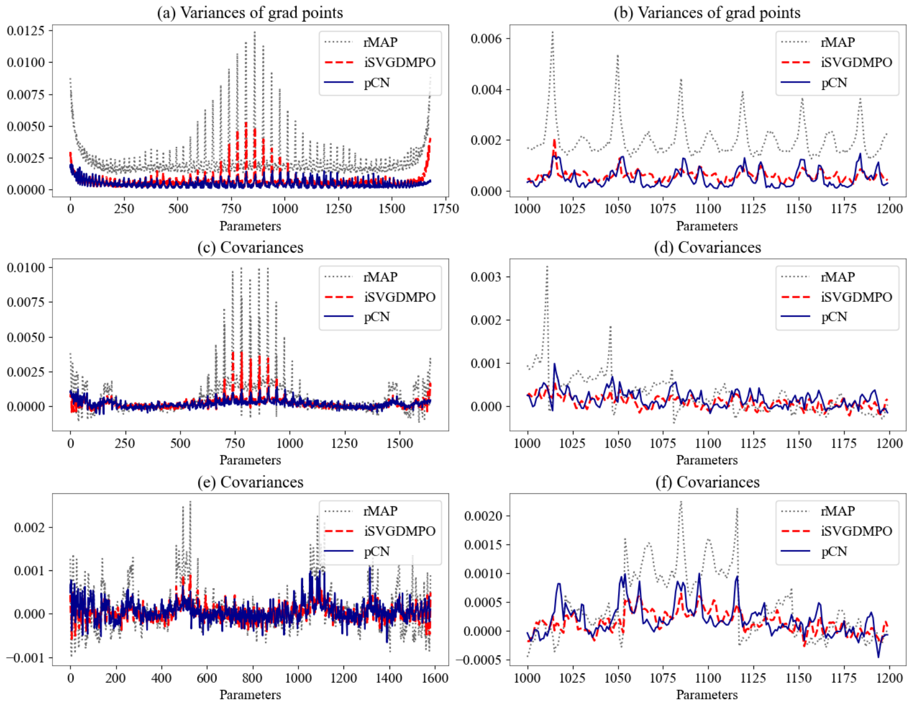

FIG. 4. *The estimated variances and covariances by the pCN (blue solid line), rMAP (gray dotted line), and iSVGDMPO (red dashed line).* (a)*: The estimated variances* $\{var_u(x_i)\}_{i=1}^{N_g}$ *on all mesh points;* (b)*: The estimated variances for mesh points with indexes from* 1000 *to* 1200 *(show details);* (c)*: The estimated covariances* $\{cov_u(x_i, x_{i+50})\}_{i=1}^{N_g-50}$ *on mesh point pairs* $\{(x_i, x_{i+50})\}_{i=1}^{N_g-50}$*;* (d)*: The estimated covariances shown in* (c) *with indexes from* 1000 *to* 1200 *(show details);* (e)*: The estimated covariances* $\{cov_u(x_i, x_{i+100})\}_{i=1}^{N_g-100}$ *on mesh point pairs* $\{(x_i, x_{i+100})\}_{i=1}^{N_g-100}$*;* (f)*: The estimated covariances shown in* (e) *with indexes from* 1000 *to* 1200 *(show details).*

TABLE 1
The ℓ^2*-norm error of the variance and covariance functions on mesh points for the rMAP and iSVGDMPO (the estimates of the pCN are seen as the background truth).*

	$\mathrm{var}_u(x_i)$	$\mathrm{cov}_u(x_i, x_{i+10})$	$\mathrm{cov}_u(x_i, x_{i+20})$	$\mathrm{cov}_u(x_i, x_{i+30})$
rMAP	0.00759	0.00100	0.00075	0.00092
iSVGDMPO	0.00038	0.00012	0.00009	0.00010
	$\mathrm{cov}_u(x_i, x_{i+40})$	$\mathrm{cov}_u(x_i, x_{i+50})$	$\mathrm{cov}_u(x_i, x_{i+60})$	$\mathrm{cov}_u(x_i, x_{i+70})$
rMAP	0.00227	0.00038	0.00043	0.00056
iSVGDMPO	0.00015	0.00007	0.00006	0.00007
	$\mathrm{cov}_u(x_i, x_{i+80})$	$\mathrm{cov}_u(x_i, x_{i+90})$	$\mathrm{cov}_u(x_i, x_{i+100})$	$\mathrm{cov}_u(x_i, x_{i+110})$
rMAP	0.00142	0.00029	0.00031	0.00047
iSVGDMPO	0.00012	0.00006	0.00006	0.00007

are the ℓ^2 differences between the vectors obtained by the rMAP and iSVGDMPO with the pCN, respectively. All of the ℓ^2 differences of the iSVGDMPO with the pCN are much smaller than the corresponding values of rMAP, which show the superiority of the iSVGDMPO.

5. Conclusion.

In this paper, the approximate sampling algorithm is proposed for the infinite-dimensional Bayesian approach. We introduce the Stein operator on Hilbert spaces and show that it is the limit of a particular finite-dimensional version. Besides, we construct the update perturbation of the SVGD on infinite-dimensional space (called iSVGD) by using the properties of operator-valued RKHS. To accelerate the convergence speed of iSVGD, we investigate the change of variables formula and introduced preconditioning operators. As examples, we present the fixed preconditioning operators and mixture preconditioning operators. Then, we calculate the explicit form of the update directions for the iSVGD with iSVGDMPOs. Finally, we apply the constructed algorithms to an inverse problem of the steady state Darcy flow equation. Comparing with the pCN and rMAP sampling algorithms, we demonstrate by numerical experiments that the proposed algorithms can generate accurate estimates efficiently.

The iSVGD is analyzed by studying the limiting behavior of the finite-dimensional objects. This work presents an infinite-dimensional version of the approach given in [58]. It is worth mentioning that our results not only provide an infinite-dimensional version but also indicate that an intuitive trivial generalization of algorithms given in [58] may not be suitable since particles will belong to a set with zero measure. Our results also show that it is necessary to introduce the parameter s, which has not been considered in the existing work.

The current work may be extended to combine the generalizations of the kernel using Hessian operators in the Wasserstein space [36]. The proposed approach may be combined with other algorithms, such as the accelerated information gradient flows [60] and the mean-field type MCMC algorithms [22], to generate new and more efficient algorithms. It is also interesting and important to do more theoretical studies, e.g., introduce infinite-dimensional Stein geometry [33] and develop systematic theories of the interacting particle system and the mean-field limit equation [42]. We will report the progress on these aspects elsewhere in the future.

Acknowledgment. The authors would like to thank Professor Wei Gong for providing helpful discussions and the anonymous referees for their comments and suggestions, which helped improve the paper quality significantly.

REFERENCES

[1] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb, *Solving inverse problems using data-driven models*, Acta Numer., 28 (2019), pp. 1–174.

[2] A. Beskos, A. Jasra, E. A. Muzaffer, and A. M. Stuart, *Sequential Monte Carlo methods for Bayesian elliptic inverse problems*, Stat. Comput., 25 (2015), pp. 727–737.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[4] T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler, *A computational framework for infinite-dimensional Bayesian inverse problems Part* 1*: The linearized case, with application to global seismic inversion*, SIAM J. Sci. Comput., 35 (2013), pp. A2494–A2523.

[5] M. Burger and F. Lucka, *Maximum a posteriori estimates in linear inverse problems with log-concave priors are proper Bayes estimators*, Inverse Problems, 30 (2014), 114004.

[6] T. But-Thanh and Q. P. Nguyen, *FEM-based discretization-invariant MCMC methods for PDE-constrained Bayesian inverse problems*, Inverse Probl. Imaging, 10 (2016), pp. 943–975.

[7] C. Carmeli, E. D. Vito, and A. Toigo, *Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem*, Anal. Appl., 4 (2006), pp. 377–408.

[8] C. Carmeli, E. D. Vito, and A. Toigo, *Vector-valued reproducing kernel Hilbert spaces and universality*, Anal. Appl., 8 (2010), pp. 19–61.

[9] E. D. C. Carvalho, R. Clark, A. Nicastro, and P. H. J. Kelly, *Scalable uncertainty for computer vision with functional variational inference*, in CVPR, IEEE Computer Society, Los Alamitos, CA, 2020, pp. 12003–12013.

[10] P. Chen and O. Ghattas, *Stein variational reduced basis Bayesian inversion*, SIAM J. Sci. Comput., 43 (2021), pp. A1163–A1193.

[11] P. Chen, K. Wu, J. Chen, T. O'Leary-Roseberry, and O. Ghattas, *Projected Stein variational Newton: a fast and scalable Bayesian inference method in high dimensions*, in NeurIPS, Vol. 32, Curran Associates, Red Hook, NY, 2020.

[12] S. L. Cotter, M. Dashti, J. C. Robinson, and A. M. Stuart, *Bayesian inverse problems for functions and applications to fluid mechanics*, Inverse Problems, 25 (2009), 115008.

[13] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White, *MCMC methods for functions: modifying old algorithms to make them faster*, Statist. Sci., 28 (2013), pp. 424–446.

[14] T. Cui, K. J. H. Law, and Y. M. Marzouk, *Dimension-independent likelihood-informed MCMC*, J. Comput. Phys., 304 (2016), pp. 109–137.

[15] G. DaPrato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, 1992.

[16] M. Dashti and A. M. Stuart, *The Bayesian approach to inverse problems*, in Handbook of Uncertainty Quantification, Springer, Cham, Switzerland, 2017, pp. 311–428.

[17] G. Detommaso, T. Cui, A. Spantini, and Y. Marzouk, *A Stein variational Newton method*, in NeurIPS, Vol. 32, Curran Associates, Red Hook, NY, 2020.

[18] A. Duncan, N. Nüsken, and L. Szpruch, *On the Geometry of Stein Variational Gradient Descent*. arXiv:1912.00894, 2019.

[19] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Springer, Netherlands, 1996.

[20] Z. Feng and J. Li, *An adaptive independence sampler MCMC algorithm for Bayesian inferences of functions*, SIAM J. Sci. Comput., 40 (2018), pp. A1301–A1321.

[21] A. Fichtner, *Full Seismic Waveform Modelling and Inversion*, Springer, New York, 2011.

[22] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart, *Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler*, SIAM J. Appl. Dyn. Syst., 19 (2020), pp. 412–441.

[23] N. Guha, X. Wu, Y. Efendiev, B. Jin, and B. K. Malick, *A variational Bayesian approach for inverse problems with skew-t error distribution*, J. Comput. Phys., 301 (2015), pp. 377–393.

[24] T. Helin and M. Burger, *Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems*, Inverse Problems, 31 (2015), 085009.

[25] J. Jia, J. Peng, and J. Gao, *Posterior contraction for empirical Bayesian approach to inverse problems under non-diagonal assumption*, Inverse Probl. Imaging, 15 (2020), pp. 201–228.

[26] J. Jia, B. Wu, J. Peng, and J. Gao, *Recursive linearization method for inverse medium scattering problems with complex mixture Gaussian error learning*, Inverse Problems, 35 (2019), 075003.

[27] J. Jia, S. Yue, J. Peng, and J. Gao, *Infinite-dimensional Bayesian approach for inverse scattering problems of a fractional Helmholtz equation*, J. Funct. Anal., 275 (2018), pp. 2299–2332.

[28] J. Jia, Q. Zhao, Z. Xu, D. Meng, and Y. Leung, *Variational Bayes' method for functions with applications to some inverse problems*, SIAM J. Sci. Comput., 43 (2021), pp. A355–A383.

[29] B. Jin, *A variational Bayesian method to inverse problems with implusive noise*, J. Comput. Phys., 231 (2012), pp. 423–435.

[30] B. Jin and J. Zou, *Hierarchical Bayesian inference for ill-posed problems via variational method*, J. Comput. Phys., 229 (2010), pp. 7317–7343.

[31] H. Kadri, E. Duflos, P. Preus, S. Canu, A. Rakotomamonjy, and J. Audiffren, *Operator-valued kernels for learning from functional response data*, J. Mach. Learn. Res., 17 (2016), pp. 1–54.

[32] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer, New York, 2005.

[33] A. Korba, A. Salim, M. Arbel, G. Luise, and A. Gretton, *A non-asymptotic analysis for Stein variational gradient descent*, in NeurIPS, Vol. 33, Curran Associates, New York, 2020.

[34] J. Lei, *Convergence and concentraction of empirical measures under Wasserstein distance in unbounded functional space*, Bernoulli, 26 (2020), pp. 767–798.

[35] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*, 2nd ed., American Mathematical Society, Providence, RI, 2017.

[36] W. C. Li, *Hessian metric via transport information geometry*, J. Math. Phys, 62 (2021), 033301.

[37] C. Liu, J. Zhuo, P. Cheng, R. Zhang, and J. Zhu, *Understanding and accelerating particle-based variational inference*, in ICML, Vol. 97, Curran Associates, Red Hook, NY, 2019, pp. 4082–4092.

[38] Q. Liu, *Stein variational gradient descent as gradient flow*, in NeurIPS, Vol. 30, Neural Information Processing Systems Foundation, La Jolla, CA, 2016.

[39] Q. Liu and D. Wang, *Stein variational gradient descent: A general purpose Bayesian inference algorithm*, in NeurIPS, Vol. 29, Neural Information Processing Systems Foundation, La Jolla, CA, 2016.

[40] A. Logg, K. A. Mardal, and G. N. Wells, *Automated Solution of Differential Equations by the Finite Element Method*, Springer, Berlin, 2012.

[41] J. C. D. los Reyes, *Numerical PDE-Constrained Optimization*, Springer, New York, 2015.

[42] J. Lu, Y. Lu, and J. Nolen, *Scaling limit of the Stein variational gradient descent: The mean field regime*, SIAM J. Math. Anal., 5 (2019), pp. 648–671.

[43] A. G. D. G. Matthews, *Scalable Gaussian Process Inference Using Variational Methods*, PhD thesis, University of Cambridge, Cambridge, 2016.

[44] R. Nickl, *Bernstein-von Mises theorem for statistical inverse problems I: Schrödinger equation*, J. Eur. Math. Soc. (JEMS), 22 (2020), pp. 2697–2750.

[45] F. J. Pinski, G. Simpson, A. M. Stuart, and H. Weber, *Algorithms for Kullback-Leibler approximation of probability measures in infinite dimensions*, SIAM J. Sci. Comput., 37 (2015), pp. A2733–A2757.

[46] F. J. Pinski, G. Simpson, A. M. Stuart, and H. Weber, *Kullback-Leibler approximation for probability measures on infinite dimensional space*, SIAM J. Math. Anal., 47 (2015), pp. 4091–4122.

[47] G. D. Prato, *Kolmogorov Equations for Stochastic PDEs*, Birkhäuser, Basel, 2004.

[48] G. D. Prato, *An Introduction to Infinite-Dimensional Analysis*, Springer, Berlin, 2006.

[49] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, 2nd ed., Springer, New York, 2005.

[50] M. Reed and B. Simon, *Functional Analysis I: Methods of Modern Mathematical Physics*, Academic Press, New York, 2003.

[51] A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. Marzouk, *Optimal low-rank approximations of Bayesian linear inverse problems*, SIAM J. Sci. Comput., 37 (2015), pp. A2451–A2487.

[52] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer, Germany, 2006.

[53] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.

[54] S. Sun, G. Zhang, J. Shi, and R. Grosse, *Functional variational Bayesian neural networks*, in ICLR, International Machine Learning Society, San Diego, CA, 2019.

[55] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, United States, 2005.

[56] A. Tarantola and B. Valette, *Inverse problems = quest for information*, J. Geophys., 50 (1982), pp. 159–170.

[57] N. G. Trillos and D. Slepčev, *On the rate of convergence of empirical measures in ∞-transportation distance*, Canad. J. Math., 67 (2015), pp. 1358–1383.

[58] D. Wang, Z. Tang, C. Bajaj, and Q. Liu, *Stein variational gradient descent with matrix-valued kernels*, in NeurIPS, vol. 33, 2019.

[59] K. Wang, T. Bui-Thanh, and O. Ghattas, *A randomized maximum a posteriori method for posterior sampling of high dimensional nonlinear Bayesian inverse problems*, SIAM J. Sci. Comput., 40 (2018), pp. A142–A171.

[60] Y. Wang and W. C. Li, *Accelerated information gradient flows*, J. Sci. Comput., 90 (2022), 11.

[61] Z. Wang, T. Ren, J. Zhu, and B. Zhang, *Function space particle optimization for Bayesian neural networks*, in ICLR, International Machine Learning Society, San Diego, CA, 2019.

[62] C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt, *Advances in variational inference*, IEEE Trans. Pattern Anal., 41 (2018), pp. 2008–2026.

[63] Q. Zhao, D. Meng, Z. Xu, W. Zuo, and Y. Yan, *l_1-norm low-rank matrix factorization by variational Bayesian method*, IEEE Trans. Neural. Netw. Learn. Syst., 26 (2015), pp. 825–839.

[64] Q. Zhou, T. Yu, X. Zhang, and J. Li, *Bayesian inference and uncertainty quantification for medical image reconstruction with Poisson data*, SIAM J. Imaging Sci., 13 (2020), pp. 29–52.