Multilevel Stein variational gradient descent with applications to Bayesian inverse problems

Terrence Alsup* Luca Venturi* Benjamin Peherstorfer*

April 2021

This work presents a multilevel variant of Stein variational gradient descent to more efficiently sample from target distributions. The key ingredient is a sequence of distributions with growing fidelity and costs that converges to the target distribution of interest. For example, such a sequence of distributions is given by a hierarchy of ever finer discretization levels of the forward model in Bayesian inverse problems. The proposed multilevel Stein variational gradient descent moves most of the iterations to lower, cheaper levels with the aim of requiring only a few iterations on the higher, more expensive levels when compared to the traditional, single-level Stein variational gradient descent variant that uses the highest-level distribution only. Under certain assumptions, in the mean-field limit, the error of the proposed multilevel Stein method decays by a log factor faster than the error of the single-level counterpart with respect to computational costs. Numerical experiments with Bayesian inverse problems show speedups of more than one order of magnitude of the proposed multilevel Stein method compared to the single-level variant that uses the highest level only.

Keywords: Monte Carlo, multilevel and multifidelity, particle methods, Bayesian inference

1. Introduction

Sampling from a target distribution \( \pi \) is a common task in Bayesian inference. Typically, in machine learning, the (unnormalized) density of the target distribution can be evaluated to approximately sample from it with Monte Carlo, variational, and particle methods \cite{39, 37, 38, 43}. We look at a setup that is more common in scientific machine learning and scientific computing, where a sequence of distributions \( (\pi(\ell)) \) is given that converges weakly to a computationally intractable target \( \pi \) for increasing level \( \ell \to \infty \). Here, intractable means that one cannot numerically evaluate the (unnormalized) density of \( \pi \). For example, one finds such a setup in Bayesian inverse problems \cite{40, 24, 30}, where the target \( \pi \) corresponds to a posterior distribution that depends on a forward model through the likelihood. The forward model is typically a system of partial differential equations (PDEs) for which only numerical solutions can be computed; increasingly more accurate, more expensive discretizations (e.g., mesh width going to 0) of the forward-model PDEs then give rise to a sequence of distributions \( (\pi(\ell)) \) that converges to \( \pi \).

To approximately sample from the target \( \pi \), one then selects a level \( L \) such that \( \pi(L) \) is a sufficiently accurate approximation of \( \pi \) and then applies Monte Carlo or particle methods to \( \pi(L) \); see, e.g., \cite{40, 24, 30}. One challenge of such an approach is that the density of \( \pi(L) \) can be computationally expensive to evaluate, because each evaluation of the density entails at least one numerical solve of the PDEs underlying the forward model, which can quickly make sampling from it prohibitively expensive.

*Courant Institute of Mathematical Sciences, New York University (alsup@cims.nyu.edu, venturi@cims.nyu.edu, pehersto@cims.nyu.edu)
Our contributions We propose to extend Stein variational gradient descent (SVGD) [29] to a multilevel SVGD (MLSVGD) that leverages the distributions from all levels \( \ell = 1, \ldots, L \) to more efficiently approximately sample from \( \pi \) than traditional, single-level SVGD that uses the distribution \( \pi^{(L)} \) on the highest level \( L \) only; thus, the proposed MLSVGD builds on the long history of exploiting hierarchies of discretizations in scientific computing (see below for literature review). Our contributions are as follows: (1) an analysis that shows the cost complexity of the proposed MLSVGD is lower than the cost complexity of single-level SVGD; (2) a numerical algorithm that builds on an adaptive stopping criterion that can be applied in a black-box way; (3) numerical experiments with Bayesian inverse problems involving nonlinear diffusion-reaction and Euler-Bernoulli beam models that demonstrate that taking into account all levels \( \ell = 1, \ldots, L \) can lead to more than one order of magnitude speedup compared to single-level SVGD.

Related work on multilevel methods in scientific computing Taking into account various discretizations and approximations of forward models to achieve computational speedups has a long tradition in scientific computing, e.g., multigrid solvers [18, 4], sparse grid approximations [5], multilevel Monte Carlo for estimating statistics [20, 14, 9]; and multifidelity methods that leverage low-fidelity models without clear hierarchies [30]. In terms of sampling from distributions, there is work on Markov chain Monte Carlo (MCMC) methods that exploit hierarchies of distributions such as multistage MCMC methods [8, 13], multilevel Metropolis–Hastings [11]; and MCMC methods with importance sampling [21]. There are multilevel/multifidelity variational methods, where a transport map (flow) is parametrized a priori; for example, [11, 35] build on [32, 33] and construct the transport maps from a distribution on a lower level and then use it as proposal for Metropolis-Hastings or for importance sampling. There are multilevel particle filters [23] and multilevel sequential Monte Carlo [5] methods, ensemble Kalman filtering [22], and extensions to nonlinear filtering using transport [15]; these rely on telescoping sums of correlated differences between successive levels, whereas our approach uses the successive levels as preconditioners for sampling. Probably closest in style to our approach are the multilevel sequential Monte Carlo method [26, 42] and the multilevel cross-entropy method [34] that use distributions obtained on lower levels as starting distributions on higher levels.

Related work on SVGD from machine learning The MLSVGD proposed in this work builds on SVGD introduced by [29] and further theoretically analyzed in [28]; extended to consider Newton directions [10]; exploiting geometry [6], and other acceleration techniques [27]. A key building block for us will be recent advances on understanding the convergence properties of SVGD in the infinite particle (mean-field) regime. The work [28, 12] shows the mean-field limit. The work [25] shows non-asymptotic results. Further, the work [7] establishes exponential convergence under certain situations in the mean-field limit that motivates some of our assumptions. Another key building block is relating discretization error of the forward model at level \( \ell \) to divergence of the corresponding posterior distributions \( \pi^{(\ell)} \) with respect to the intractable target \( \pi \), where we build on results by [40] and an inequality involving the Kullback-Leibler (KL) divergence introduced by [31].

2. Preliminaries: Approximating measures with SVGD

Let \( \Theta \subset \mathbb{R}^d \) and \( (\Theta, \mathcal{B}(\Theta)) \) be a measurable space with \( \mathcal{B}(\Theta) \) denoting the Borel \( \sigma \)-algebra of \( \Theta \). Consider approximating some target measure \( \eta \) on \( \Theta \) via an empirical measure, i.e., an ensemble of samples (particles); in the following, the distribution \( \eta_t \), and all other distributions that will be considered, admit a density with respect to the Lebesgue measure over \( \Theta \). Moreover, the target distribution has the form \( \eta \propto e^{-V} \) with the potential \( V \).

2.1. Approximating measures with SVGD

The SVGD method [29] iteratively moves forward an empirical distribution given by an ensemble \( \{\theta_t^{(i)}\}_{i=1}^N \) from time \( t \) to time \( t + \delta \) via a map \( \phi_t(\theta) = \theta - \delta g_t(\theta) \), where \( \delta \) is a step size and \( g_t : \Theta \to \mathbb{R}^d \) is a vector field. SVGD chooses \( g_t \) from a vector-valued reproducing kernel Hilbert space (RKHS) \( \mathcal{H}^d \) with kernel \( K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \) via a functional gradient descent step on the KL divergence (cf. [47] in Appendix A). Denote the distribution of the particles at time \( t \) as \( \mu_t \) and define the functional \( J_t(g) = \text{KL}( (I - g)\#\mu_t | \eta ) \), where \( (I - g)\#\mu_t \) denotes the pushforward measure.
Remark 1. Although we use the Hellinger distance $d_{\text{Hell}}$, we will see that the integration time is independent of SVGD, and (2) the costs of SVGD to find the (unnormalized) density of the ensemble of particles $\{\theta_i^{[\ell]}\}_{i=1}^N$ and the ODE is integrated using the forward Euler method. Thus, the SVGD update becomes

$$
\theta_{t+\delta}^{[\ell]} = \theta_t^{[\ell]} + \frac{\delta}{N} \left( \sum_{j=1}^N \nabla_1 K(\theta_j^{[\ell]}, \theta_t^{[\ell]}) + \sum_{j=1}^N K(\theta_j^{[\ell]}, \theta_t^{[\ell]}) \nabla \log \eta(\theta_j^{[\ell]}) \right),
$$

In [28], the distribution of the particles $\{\theta_i^{[\ell]}\}_{i=1}^N$ in the limit as $N \to \infty$ is given by the mean-field PDE

$$
\partial_t \mu_t(\theta) = -\nabla \cdot (\mu_t(\theta) \nabla \log \eta(\theta)) + \nabla_1 K(\theta, \theta),
$$

with an initial measure $\mu_0$; see also [7, 19, 28].

2.2. Approximating intractable target measures with SVGD

Consider now an intractable target distribution $\pi$; in contrast to the measure $\eta$ in Section 2.1, we cannot evaluate the (unnormalized) density of $\pi$ nor sample from $\pi$ directly. Thus, the SVGD algorithm cannot directly be applied to $\pi$. Instead, suppose we have a sequence of distributions $(\pi^{(\ell)})_{\ell \geq 1}$ that converges weakly to $\pi$ for $\ell \to \infty$ (note that weak convergence is implied by convergence in the KL divergence) and call $\ell$ the level. Moreover, we can evaluate the unnormalized density of each $\pi^{(\ell)}$ with computational costs $c_{\ell}$. Such a setup is common in Bayesian inverse problems; cf. Section 1.

Single-level approximation with SVGD. The aim is deriving a distribution $\mu$ that approximates $\pi$ with accuracy $\epsilon$. To quantify how close the approximation $\mu$ is to the target distribution, we consider the Hellinger distance $d_{\text{Hell}}(\cdot, \cdot)$ in the following: First, select a level $L \in \mathbb{N}$ such that $d_{\text{Hell}}(\pi^{(L)}, \pi) \leq \epsilon/2$. Then, to approximate $\pi^{(L)}$ with SVGD, derive $\mu$ with accuracy $d_{\text{Hell}}(\mu, \pi^{(L)}) \leq \epsilon/2$ from an initial distribution $\mu_0$; the triangle inequality leads to $d_{\text{Hell}}(\mu, \pi) \leq \epsilon$. The fact that the Hellinger distance is a metric is important because it allows us to separate the error due to truncating at level $L$ and the error due to the SVGD approximation of $\pi^{(L)}$; see Appendix A for the definition of the Hellinger distance.

Computational costs. The costs of such an approach depend on two factors: (1) the costs $c_{\ell}$ of evaluating the density $\pi^{(L)}$ on level $L$, which is independent of SVGD, and (2) the costs of SVGD to find $\mu$ from $\pi^{(L)}$ with initial distribution $\mu_0$ to achieve $d_{\text{Hell}}(\mu, \pi^{(L)}) \leq \epsilon/2$. In the continuous SVGD given by Equation (1), we identify the costs of the approximation $\mu_T$ after integrating up to end time $T$ as

$$
c_{\text{SL}}(T) = c_L T.
$$

We will see that the integration time $T$ depends on the divergence between $\mu_0$ and $\pi^{(L)}$. For the discrete SVGD given by Equation (2), time is replaced with number of iterations and the costs must be multiplied by the number of particles $N$.

**Remark 1.** Although we use the Hellinger distance $d_{\text{Hell}}$ in the following, the proposed analysis is also applicable if a different metric is used as long as it can be upper bounded by the KL divergence; see Section 3.3 for more details. Indeed, we make frequent use of the fact that the Hellinger distance can be bounded as

$$
2 d_{\text{Hell}}(\rho_1, \rho_2)^2 \leq \text{KL}(\rho_1 \parallel \rho_2)
$$
single-level SVGD:
\[ \mu_0 \xrightarrow{T} \pi^{(L)} \xrightarrow{T} \mu^{SL} \]

proposed MLSVGD:
\[ \mu_0 \xrightarrow{T_1} \mu^{(1)} T_1 \xrightarrow{T_2} \mu^{(2)} T_2 \xrightarrow{T_3} \cdots \xrightarrow{T_L} \mu^{ML} \]

Figure 1.: The proposed MLSVGD leverages a hierarchy of distributions with increasing costs and fidelity with the aim of requiring fewer iterations on the higher, more expensive levels compared to traditional, single-level SVGD that uses the highest-level distribution only.

for two distributions \( \rho_1, \rho_2 \); see Lemma 2.4 of [41] (note that the definition of Hellinger distance there is scaled by a constant factor \( \sqrt{2} \)). The Hellinger distance is also useful because it can be used to bound the bias of a Monte Carlo estimator as shown in [40].

3. A continuous multilevel Stein variational method and its cost complexity

We propose MLSVGD that leverages the measures \( \pi^{(1)}, \ldots, \pi^{(L-1)} \) with the aim to reduce the costs of approximating \( \pi^{(L)} \) compared to the traditional, single-level SVGD that uses \( \pi^{(L)} \) only. Our analysis of the proposed MLSVGD method is conducted in the time-continuous and mean-field setting where the SVGD measures satisfy the PDE (3) and the particles satisfy the ODE (1). A discrete, heuristic, algorithmic formulation follows in Section 5 with a numerical comparison to single-level SVGD in Section 6.

3.1. Continuous MLSVGD

To describe the proposed MLSVGD, consider the levels \( \ell = 1, \ldots, L \) and let \( \mu_0 \) be an initial distribution. At level \( \ell = 1 \), we define \( \mu^{(1)} T_1 \) as the distribution of the continuous SVGD (3) at time \( T_1 \) with target \( \pi^{(1)} \) and initial \( \mu_0 \). At level \( \ell = 2 \), we obtain \( \mu^{(2)} T_2 \) at time \( T_2 \) with the target \( \pi^{(2)} \) and initial distribution \( \mu^{(1)} T_1 \). In general, at level \( \ell \), we obtain \( \mu^{(\ell)} T_{\ell} \) at time \( T_{\ell} \) with target \( \pi^{(\ell)} \) and initial distribution \( \mu^{(\ell-1)} T_{\ell-1} \). Thus, deriving \( \mu^{(\ell)} T_{\ell} \) is an iterative process over the levels \( 1, \ldots, \ell - 1 \), depicted in Figure 1, of first computing \( \mu^{(1)} T_1, \ldots, \mu^{(\ell-1)} T_{\ell-1} \). The costs of MLSVGD are given by

\[ c_{ML}(T_1, \ldots, T_L) = \sum_{\ell=1}^{L} c_{\ell} T_{\ell}, \]

cf. the costs \( c_{SL}(T) = c_L T \) of the single-level SVGD as defined in (4).

3.2. Assumptions for cost complexity analysis of single-level SVGD and MLSVGD

We build on the following three assumptions to derive the cost complexity of both traditional single-level SVGD as well as the proposed MLSVGD. The first assumption is a standard assumption in scientific computing on the cost of evaluating the densities, while the second and third are needed to certify that \( d_{\text{Hell}}(\mu, \pi) \leq \epsilon \).

**Assumption 1.** The costs \( c_{\ell} \) of evaluating the (unnormalized) density \( \pi^{(\ell)} \) are bounded as

\[ c_{\ell} \leq c_0 s^{\gamma_{\ell}}, \quad \ell \in \mathbb{N}, \]
with constants $c_0, \gamma > 0$ independent of $\ell$ and $s > 1$.

**Assumption 2.** There exists $\alpha, k_0, k_1 > 0$ independent of $\ell$ such that $\text{KL}(\mu_0||\pi^{(\ell)}) \leq k_0$ for all $\ell \in \mathbb{N}$ and

$$\text{KL}(\pi^{(\ell)}||\pi) \leq k_1 s^{-\alpha\ell}, \quad \ell \in \mathbb{N},$$

where $s$ is the same constant independent of $\ell$ as in Assumption 1 and $\mu_0$ is the initial distribution.

**Assumption 3.** There exists a rate $\lambda > 0$ such that for any initial distribution $\nu_0$

$$\text{KL}(\nu_t||\pi^{(\ell)}) \leq e^{-\lambda t} \text{KL}(\nu_0||\pi^{(\ell)}), \quad \ell \in \mathbb{N},$$

holds, where $\nu_t$ solves the mean-field SVGD equation (3) at time $t$.

[25] show that Assumption 3 is satisfied if the measures $\pi^{(\ell)}$ satisfy a Stein log-Sobolev inequality. [7] also show that Assumption 3 is satisfied for a specific choice of the kernel $K$. We also note that the exponential convergence rate for the KL divergence appears in the theory for the convergence of Markov processes when the target measure satisfies a log-Sobolev inequality [2, Theorem 5.2.1]; however, SVGD approximates the gradient in an RKHS and thus [2, Theorem 5.2.1] is not directly applicable.

### 3.3. Cost complexity of continuous single-level SVGD

Consider the single-level SVGD that selects $L$ such that $d_{\text{Hell}}(\pi^{(L)}, \pi) \leq \epsilon/2$ and then starts with a $\mu_0$ to find $\mu_{SL}^T$ that satisfies $d_{\text{Hell}}(\mu_{SL}^T, \pi^{(L)}) \leq \epsilon/2$. For brevity, we write $\mu_{SL} = \mu_{SL}^T$. The following proposition bounds the costs of this single-level SVGD with respect to the tolerance $\epsilon$.

**Proposition 1.** If Assumptions 1–3 hold, then the costs of continuous single-level SVGD to obtain $\mu_{SL}$ with $d_{\text{Hell}}(\mu_{SL}, \pi) \leq \epsilon$ is bounded as

$$c_{\text{SL}}^*(\epsilon) \leq \frac{2c_0 s^{\gamma}}{\lambda} \left( \frac{\sqrt{2k_1}}{\epsilon} \right)^{2\gamma/\alpha} \log \left( \frac{\sqrt{\text{KL}(\mu_0||\pi^{(L)})}}{\sqrt{2\epsilon}} \right).$$

(7)

**Proof.** By the triangle inequality for the Hellinger distance we have that

$$d_{\text{Hell}}(\mu_{SL}, \pi) \leq d_{\text{Hell}}(\mu_{SL}, \pi^{(L)}) + d_{\text{Hell}}(\pi^{(L)}, \pi),$$

so we will bound both of these terms independently by $\epsilon/2$. By inequality (5), it is sufficient to bound the KL divergence because

$$d_{\text{Hell}}(\mu_{SL}, \pi^{(L)}) \leq \sqrt{\frac{\text{KL}(\mu_{SL}||\pi^{(L)})}{2}},$$

(8)

and similarly for $d_{\text{Hell}}(\pi^{(L)}, \pi)$. By Assumption 2 choose $L$ to be

$$L = \left[ \frac{1}{\lambda} \log_s \left( \frac{2k_1}{\epsilon^2} \right) \right] \leq \frac{1}{\alpha} \log_s \left( \frac{2k_1}{\epsilon^2} \right) + 1,$$

(9)

so that

$$d_{\text{Hell}}(\pi^{(L)}, \pi) \leq \sqrt{\frac{\text{KL}(\pi^{(L)}||\pi)}{2}} \leq \sqrt{\frac{k_1 s^{-\alpha L}}{2}} \leq \frac{\epsilon}{2}.$$  

(10)

Now by Assumptions 3 the time needed to integrate with SVGD to achieve $d_{\text{Hell}}(\mu_{SL}, \pi^{(L)}) \leq \epsilon/2$ is

$$T_{\text{SL}}^* \leq \frac{1}{\lambda} \log \left( \frac{\text{KL}(\mu_0||\pi^{(L)})}{2\epsilon^2} \right).$$

(11)
The total cost to integrate until time \( T^*_L \) at level \( L \) is thus

\[
c^*_L(\epsilon) = c_0 s^{\gamma L} T^*_L \leq \frac{2c_0 s^{\gamma}}{\lambda} \left( \frac{\sqrt{2k_1}}{\epsilon} \right)^{2\gamma/\alpha} \log \left( \frac{\sqrt{KL(\mu_0 || \pi(L))}}{\sqrt{2}\epsilon} \right).
\]

**Discussion of cost complexity of single-level SVGD**

The bound (4) in Proposition 2 shows that if we start with an initial distribution \( \mu_0 \) that has a large KL divergence \( KL(\mu_0 || \pi(L)) \) with respect to \( \pi(L) \), then we will need to integrate for a long time with SVGD to reach our tolerance. The proposed MLSVGD is aiming to avoid the long time integration by starting the integration at the highest level \( L \) with good initial distributions found on the cheaper, lower levels \( \ell = 1, \ldots, L - 1 \) that are closer to \( \pi(L) \) in the KL divergence than \( \mu_0 \).

### 3.4. Cost complexity of continuous MLSVGD

Consider now the MLSVGD approach of Section 3.1. We need to make one additional assumption compared to the single-level SVGD regarding the KL divergence between consecutive measures \( \pi(\ell) \) and \( \pi(\ell - 1) \) that will allow us to chain them together as in Figure 1.

**Assumption 4.** There exists a constant \( k_2 > 0 \) independent of \( \ell \) such that \( KL(\pi(\ell - 1)||\pi(\ell)) \leq k_2 s^{-\alpha \ell} \), where \( \alpha \) is the same rate as in Assumption 2.

The key result is to use a triangle-like inequality as in Appendix D to decompose the KL divergence. In particular,

\[
KL\left( \mu_{T_\ell}^{(\ell-1)} || \pi(\ell) \right) = KL\left( \mu_{T_{\ell-1}}^{(\ell-1)} || \pi(\ell-1) \right) + KL\left( \pi(\ell-1) || \pi(\ell) \right) + R_\ell
\]

with the remainder \( R_\ell \) given by

\[
R_\ell = \int_{\mathbb{R}^d} \left( \mu_{T_{\ell-1}}^{(\ell-1)}(\theta) - \pi(\ell-1)(\theta) \right) \log \left( \frac{\pi(\ell-1)(\theta)}{\pi(\ell)(\theta)} \right) d\theta.
\]

Because \( \pi(\ell) \) converges to \( \pi \), we have that \( \pi(\ell-1)/\pi(\ell) \to 1 \) pointwise and hence \( \log (\pi(\ell-1)/\pi(\ell)) \to 0 \). Moreover, \( \mu_{T_\ell}^{(\ell)} \to \pi(\ell) \) as \( T_\ell \to \infty \). Thus, \( R_\ell \to 0 \). In particular \( R_\ell \) is a bounded sequence meaning that there is some constant \( R \geq R_\ell \) for all \( \ell \). The following proposition give bounds on the costs of MLSVGD. The later Proposition 3 will give a faster decaying bound on the costs if \( R_\ell \) goes to zero with a known rate, as in our Bayesian inverse problems in Section 4.

**Proposition 2.** If Assumptions 1, 4 hold, then continuous MLSVGD gives \( \mu^{ML} \) with \( d_{Hell}(\mu^{ML}, \pi) \leq \epsilon \) with costs bounded as

\[
c^*_{ML}(\epsilon) \leq \frac{2c_0 s^{2\gamma}}{\lambda \gamma \log(s)} \left( \frac{\sqrt{2k_1}}{\epsilon} \right)^{2\gamma/\alpha} \log \left( \frac{\sqrt{e^{2} + 2(k_2 + R)}}{\epsilon} \right),
\]

where \( R \) bounds (13).

**Proof.** As in Equation 9 in the proof of Proposition 1 we select the level \( L \) as

\[
L = \left\lceil \frac{\log(s)}{\alpha} \left( \frac{2k_1}{2^2} \right) \right\rceil \leq \frac{1}{\alpha} \log_2 \left( \frac{2k_1}{\epsilon^2} \right) + 1,
\]

so that \( d_{Hell}(\pi(L), \pi) \leq \epsilon/2 \). The total cost for the continuous MLSVGD is

\[
c^*_{ML}(\epsilon) = \sum_{\ell=1}^{L} c_0 s^{\gamma \ell} T_\ell,
\]

\[
\square
\]
where it remains to choose the integration times \( T_\ell \) at each level. To do this we balance the KL divergence of the SVGD approximation with the KL divergence due to the fidelity. By Equation (12), we have

\[
\text{KL}(\mu_{\ell}^{(\ell)} || \pi^{(\ell)}) \leq e^{-\lambda T_\ell \left( \text{KL}(\mu_{\ell-1}^{(\ell-1)} || \pi^{(\ell-1)}) + \text{KL}(\pi^{(\ell-1)} || \pi^{(\ell)}) + R_\ell \right)},
\]

giving a recursive bound on the KL divergence in terms of the KL divergence at the previous level. At each level \( \ell \) choose the integration time \( T_\ell \) so that

\[
\text{KL}(\mu_{\ell}^{(\ell)} || \pi^{(\ell)}) \leq \frac{\epsilon^2}{2}
\]

is satisfied. In particular, at the final level \( L \) we will have that \( \text{KL}(\mu^{ML} || \pi^{(L)}) \leq \epsilon^2/2 \) and hence \( d_{\text{Hell}}(\mu^{ML} || \pi^{(L)}) \leq \epsilon/2 \) as desired. By choosing \( T_\ell \) so that this is satisfied at every level we have from Equation (16) that

\[
\text{KL}(\mu_{\ell}^{(\ell)} || \pi^{(\ell)}) \leq e^{-\lambda T_\ell \left( \frac{\epsilon^2}{2} + \text{KL}(\pi^{(\ell-1)} || \pi^{(\ell)}) + R_\ell \right)} \leq \frac{\epsilon^2}{2}.
\]

Thus, we choose \( T_\ell \) sequentially so that Equation (17) is always satisfied. As a result, the integration time \( T_\ell \) needed at each level \( \ell \) is bounded by

\[
T_\ell \leq \frac{1}{\lambda} \log \left( 1 + 2\left( \frac{\text{KL}(\pi^{(\ell-1)} || \pi^{(\ell)}) + R_\ell}{\epsilon^2} \right) \right).
\]

Finally, the total cost can be bounded by

\[
c_{ML}^*(\epsilon) \leq \sum_{\ell=1}^{L} \frac{c_0}{\lambda} s^{\gamma \ell} \log \left( 1 + 2\left( \frac{\epsilon^2}{2} + \text{KL}(\pi^{(\ell-1)} || \pi^{(\ell)}) + R_\ell \right) \right).
\]

We now use the fact that \( R_\ell \leq R \) and \( \text{KL}(\pi^{(\ell-1)} || \pi^{(\ell)}) \leq k_2 \) to obtain

\[
c_{ML}^*(\epsilon) \leq \sum_{\ell=1}^{L} \frac{c_0}{\lambda} s^{\gamma \ell} \log \left( 1 + 2\left( \frac{k_2 + R}{\epsilon^2} \right) \right).
\]

Since the terms in this sum are increasing, we can upper bound the cost further by switching to an integral

\[
c_{ML}^*(\epsilon) \leq \int_0^{L+1} \frac{c_0}{\lambda} s^{\gamma x} \log \left( 1 + 2\left( \frac{k_2 + R}{\epsilon^2} \right) \right) \, dx
\]

\[
= \frac{c_0}{\lambda \gamma \log(s)} \log \left( 1 + 2\left( \frac{k_2 + R}{\epsilon^2} \right) \right) s^{\gamma(L+1)}
\]

\[
\leq \frac{2c_0 s^{2\gamma}}{\lambda \gamma \log(s)} \left( \frac{2k_1}{\epsilon} \right)^{2\gamma/\alpha} \log \left( \frac{\sqrt{\epsilon^2 + 2(k_2 + R)}}{\epsilon} \right).
\]

We now consider the case where the remainder term behaves as \( R_\ell \lesssim s^{-\alpha \ell} \), which allows us to make a more efficient choice when selecting the integration time \( T_\ell \) at each level. In particular, it allows us to set \( T_\ell \) such that

\[
\text{KL}(\mu_{\ell}^{(\ell)} || \pi^{(\ell)}) \sim s^{-\alpha \ell},
\]

which leads to the following proposition that shows an improved cost complexity compared to Proposition 2.
Proposition 3. If Assumptions 1–4 hold and \( R_L \leq k_3 s^{-\alpha} \), then the costs of continuous MLSVGD to have \( d_{\text{Hel}}(\mu^{\text{ML}}, \pi) \leq \epsilon \) can be bounded as

\[
\tilde{c}^*_{\text{ML}}(\epsilon) \leq \frac{c_0 s^{2\gamma}}{\lambda \gamma \log(s)} \log \left( s^\alpha + \frac{k_2 + k_3}{k_1} \right) \left( \frac{\sqrt{2k_1}}{\epsilon} \right)^{2\gamma/\alpha}.
\] (24)

Proof. Starting from Equation (16) in the proof of Proposition 2 change \( \epsilon \) to instead be

\[
\epsilon_\ell = \sqrt{2k_1 s^{-\alpha} \ell/2}
\] (25)
at each level \( \ell \). By Assumption 2 we know that \( L \) is chosen so that

\[
\epsilon^2_L = 2k_1 s^{-\alpha} L \leq \epsilon^2,
\] (26)
so that \( \epsilon_L \leq \epsilon \). Plugging in this choice gives that the integration times needed are

\[
T^*_\ell \leq \frac{1}{s^\alpha} \log \left( s^\alpha + \frac{k_2 + k_3}{k_1} \right).
\] (27)

By Assumption 4 and the assumption in the proposition, we have that

\[
T^*_\ell \leq \frac{1}{s^\alpha} \log \left( s^\alpha + \frac{k_2 + k_3}{k_1} \right),
\] (28)
so that the integration time is fixed at each level. The cost is now bounded by

\[
\tilde{c}^*_{\text{ML}}(\epsilon) \leq \sum_{\ell=1}^{L} \frac{c_0 s^{\gamma \ell}}{\lambda} \log \left( s^\alpha + \frac{k_2 + k_3}{k_1} \right).
\] (29)

Since the terms in the sum are increasing, we can further bound this with an integral:

\[
\tilde{c}^*_{\text{ML}}(\epsilon) \leq \sum_{\ell=1}^{L} \frac{c_0 s^{\gamma \ell}}{\lambda} \log \left( s^\alpha + \frac{k_2 + k_3}{k_1} \right) \leq \int_1^{L+1} \frac{c_0 s^{\gamma x}}{\lambda} \log \left( s^\alpha + \frac{k_2 + k_3}{k_1} \right) \, dx.
\] (30)

Computing the integral gives

\[
\tilde{c}^*_{\text{ML}}(\epsilon) \leq \frac{c_0 s^{2\gamma}}{\lambda \gamma \log(s)} \log \left( s^\alpha + \frac{k_2 + k_3}{k_1} \right) s^{\gamma L}.
\] (31)

Finally, by plugging in \( L \) we obtain

\[
\tilde{c}^*_{\text{ML}}(\epsilon) \leq \frac{c_0 s^{2\gamma}}{\lambda \gamma \log(s)} \log \left( s^\alpha + \frac{k_2 + k_3}{k_1} \right) \left( \frac{\sqrt{2k_1}}{\epsilon} \right)^{2\gamma/\alpha}.
\] (32)

Discussion of cost complexity of MLSVGD Looking at the single-level SVGD and MLSVGD cost bounds from Propositions 1 and 3 respectively, we note two major differences. The first is that there is no \( \log \epsilon^{-1} \) term in the cost bound (24) of Proposition 3 and thus MLSVGD achieves a cost complexity that grows by \( \log \epsilon^{-1} \) slower than the cost complexity of single-level SVGD as \( \epsilon \to 0 \). Moreover, whenever \( \epsilon \to 0 \), we now have a fixed integration time at each level \( \ell \) as opposed to requiring an increasing number of iterations as the level goes to infinity as in the single-level case. The second notable difference is that the constant \( k_0 \), which depends on the KL divergence from the initial distribution \( \mu_0 \) and the target \( \pi \), does not appear in (24). Instead the bound (24) depends on the constant \( k_2 \) from Assumption 4 which depends only on the KL divergence between two consecutive levels. Thus, if the KL divergence between consecutive levels is low, then the previous level serves as a good preconditioner for the next level leading to reduced costs.
Remark 2. The order $\log \epsilon^{-1}$ comes from the exponential decay rate of the KL divergence for SVGD in Assumption 3. If the assumption is violated and, for example, the KL divergence decays only algebraically, then we expect the speedup to be on the order of $\epsilon^{-\beta}$ for some constant $\beta > 0$. This is further supported by our numerical results that indicate that MLSVGD obtains speedups even if SVGD converges slower than in Assumption 3. We leave the detailed analysis of this to future work.

4. MLSVGD for Bayesian inverse problems

Typically, in Bayesian inverse problems in scientific computing, one is interested in inferring an unknown quantity $\theta$ from some noisy observed data $y = G(\theta^*) + e$ with $G$ denoting the parameter-to-observable map and $e$ being the noise; see, e.g., [40, 24, 30]. Let $\pi$ be the prior and consider zero-mean Gaussian noise with covariance $\Gamma$, then the posterior is given by

$$\pi(\theta) = \frac{1}{Z} \exp \left( -\frac{1}{2} \|y - G(\theta)\|_{\Gamma^{-1}}^2 \right) \pi_0(\theta),$$

with the normalizing constant

$$Z = \int_\Theta \exp \left( -\frac{1}{2} \|y - G(\theta)\|_{\Gamma^{-1}}^2 \right) \pi_0(\theta) \, d\theta,$$

where $\|u\|_{r^{-1}} = \langle \Gamma^{-1}u, u \rangle$. Now let $(G_\ell)_{\ell \geq 1}$ denote a sequence of approximations to the parameter-to-observable map $G$, e.g., given by finite-difference or finite-element discretizations of the PDEs underlying $G$, and define $\pi^{(\ell)}$ by replacing $G$ with $G_\ell$ and define $Z_\ell$ similarly. The next two assumptions will be sufficient to apply our results from Section 3.

Assumption 5 (Model error). There is a function $\psi : \mathbb{N} \to (0, \infty)$, with $\psi(\ell) \to 0$ as $\ell \to \infty$, such that

$$\|G(\theta) - G_\ell(\theta)\|_{L^2(\pi_0)} \leq \psi(\ell),$$

(35)

where the $\cdot \|_{L^2(\pi_0)}$ is the $L^2$ norm over $\pi_0$; cf. [49] in Appendix A.

Assumption 6. There exists a constant $b_3 > 0$ independent of $\ell$ such that

$$\mu^{(\ell)}(\theta) \leq b_3 \pi_0(\theta)$$

(36)

for all $\ell \geq 1$.

The next theorem shows that if Assumptions 1, 3, 5, and 6 are satisfied, then our cost complexity results derived for MLSVGD in Section 3 hold in the Bayesian inverse problem setting. These assumptions can be interpreted in the context of Bayesian inverse problems as follows: Assumption 1 and Assumption 5 are related to the forward model. Together they state that the approximation $G_\ell$ converges in an $L^2$-sense to $G$ as the level $\ell$ is increased. At the same time, as the level $\ell$ is increased and $G_\ell$ gets closer to $G$, the computational costs of evaluating $G_\ell$ may increase with a rate $\gamma$. This is typical behavior in, e.g., finite-element forward models where refining the mesh (increasing the level) leads to more accurate approximations and at the same time the computational costs of computing the finite-element solution increase with the number of mesh points. Furthermore, Assumption 5 is similar to the assumptions of [40] Corollary 4.9, although there a pointwise bound is used. Assumption 3 is the convergence rate of SVGD and motivated by results from the literature as discussed in Section 3 cf. Remark 2 for other convergence behavior. Assumption 6 ensures that the tail of the posterior distribution behaves as the tail of the prior and is similar to the envelope assumption made in, e.g., acceptance/rejection sampling [39].

Theorem 1. If Assumptions 1, 3, 5, and 6 hold and Assumption 3 holds with $\psi(\ell) = b_0 s^{-\alpha \ell}$, then Assumptions 2 and 4 hold and thus the cost complexity to find $\mu^{\text{ML}}$ with $d_{\text{Hell}}(\mu^{\text{ML}}, \pi) \leq \epsilon$ is given by

$$c^*_\text{ML}(\epsilon) \leq \frac{c_0 s^{2\gamma}}{\lambda \gamma \log(s)} \log \left( \frac{1}{\epsilon} \left( \frac{4}{3} + \frac{b_3}{3b_1 b_2} \right) \left( \frac{e}{\sqrt{3b_1 b_2 b_0}} \right)^{2\gamma/\alpha} \right),$$

(37)

where the constants $b_1, b_2$ are independent of $\epsilon$ and given in the proof of Lemma 1 in Appendix B.
Proof. By Lemma 2 in Appendix C, we know that Assumptions 2 and 4 hold with $k_1 = Cb_0$ and $k_2 = Cb_0(1 + s^\alpha)$. Thus, we just need to verify that $R_\ell \leq k_3s^{-\alpha\ell}$ for some constant $k_3$. To apply Proposition 2, we have that

$$ R_\ell = \int_\Theta \left( \mu_{T_{\ell-1}}^{(\ell-1)}(\theta) - \pi^{(\ell-1)}(\theta) \right) \log \left( \frac{\pi^{(\ell-1)}(\theta)}{\mu^{(\ell-1)}(\theta)} \right) \, d\theta $$

$$ = \int_\Theta \left( \mu_{T_{\ell-1}}^{(\ell-1)}(\theta) - \pi^{(\ell-1)}(\theta) \right) \log \left( \frac{Z_\ell \exp \left( -\frac{1}{2} \| y - G_{\ell-1}(\theta) \|^2_{T_{\ell-1}} \right)}{Z_{\ell-1} \exp \left( -\frac{1}{2} \| y - G_{\ell}(\theta) \|^2_{T_{\ell-1}} \right)} \right) \, d\theta $$

$$ = \int_\Theta \left( \mu_{T_{\ell-1}}^{(\ell-1)}(\theta) - \pi^{(\ell-1)}(\theta) \right) \log \left( \frac{\exp \left( -\frac{1}{2} \| y - G_{\ell-1}(\theta) \|^2_{T_{\ell-1}} \right)}{\exp \left( -\frac{1}{2} \| y - G_{\ell}(\theta) \|^2_{T_{\ell-1}} \right)} \right) \, d\theta, $$

where the last line follows from the fact that

$$ \int_\Theta \left( \mu_{T_{\ell-1}}^{(\ell-1)}(\theta) - \pi^{(\ell-1)}(\theta) \right) \log \left( \frac{Z_\ell}{Z_{\ell-1}} \right) \, d\theta = 0 $$

since $\frac{Z_\ell}{Z_{\ell-1}}$ is a constant and $\pi^{(\ell-1)}$ and $\mu_{T_{\ell-1}}^{(\ell-1)}$ both integrate to one. By the triangle inequality we have that

$$ R_\ell \leq \frac{1}{2} \int_\Theta \left( \| y - G_{\ell}(\theta) \|^2_{T_{\ell-1}} - \| y - G_{\ell-1}(\theta) \|^2_{T_{\ell-1}} \right) \mu_{T_{\ell-1}}^{(\ell-1)}(\theta) \, d\theta $$

$$ + \frac{1}{2} \int_\Theta \left( \| y - G_{\ell}(\theta) \|^2_{T_{\ell-1}} - \| y - G_{\ell-1}(\theta) \|^2_{T_{\ell-1}} \right) \pi^{(\ell-1)}(\theta) \, d\theta $$

We have that

$$ \pi^{(\ell-1)}(\theta) \leq \frac{1}{Z_{\ell-1}} \pi_0(\theta), $$

so that when combined with Assumption 6

$$ R_\ell \leq \frac{1}{2} \int_\Theta \left( \| y - G_{\ell}(\theta) \|^2_{T_{\ell-1}} - \| y - G_{\ell-1}(\theta) \|^2_{T_{\ell-1}} \right) \mu_{T_{\ell-1}}^{(\ell-1)}(\theta) \, d\theta $$

$$ + \frac{1}{2} \int_\Theta \left( \| y - G_{\ell}(\theta) \|^2_{T_{\ell-1}} - \| y - G_{\ell-1}(\theta) \|^2_{T_{\ell-1}} \right) \pi^{(\ell-1)}(\theta) \, d\theta $$

$$ \leq \frac{b_3}{2} \int_\Theta \left( \| y - G_{\ell}(\theta) \|^2_{T_{\ell-1}} - \| y - G_{\ell-1}(\theta) \|^2_{T_{\ell-1}} \right) \pi_0(\theta) \, d\theta $$

$$ + \frac{1}{2Z_{\ell-1}} \int_\Theta \left( \| y - G_{\ell}(\theta) \|^2_{T_{\ell-1}} - \| y - G_{\ell-1}(\theta) \|^2_{T_{\ell-1}} \right) \pi_0(\theta) \, d\theta $$

$$ \leq \left( \frac{b_3}{2} + \frac{b_1b_2}{2} \right) \| G_\ell - G_{\ell-1} \|_{L^2(\pi_0)}, $$

so that $k_3 = \left( \frac{b_3}{2} + \frac{b_1b_2}{2} \right) b_0(1 + s^\alpha)$. Plugging in the values of $k_1, k_2,$ and $k_3$ into Proposition 2 gives the result. $\square$

5. A discrete, heuristic MLSVGD algorithm with adaptive stopping criterion

In this section, we propose a discrete, heuristic MLSVGD method given in Algorithm 1 that uses an adaptive stopping criterion to decide when to switch to the next higher level. The proposed Algorithm 1 uses the estimates of the gradient norms to decide when to switch to the next higher level. Thus, the algorithm avoids requiring any constants that are not readily available in practice. In particular, the algorithm is independent of the constants and rates used in the MLSVGD cost complexity analysis to derive the optimal choice of times $T_1, \ldots, T_L$. 

10
6. Numerical experiments

We now demonstrate MLSVGD on Bayesian inverse problems: The aim is to infer the unknown coefficients of a PDE model from noisy observations of the state of the PDE at a few locations in the spatial domain. In Section 6.1, we consider a reaction-diffusion model with unknown reaction parameters, which are then inferred from measurements of the diffusion-reaction field. In Section 6.2, the displacement of an Euler-Bernoulli beam is observed and we then infer the stiffness of the beam. Details about the setup of the numerical experiments are in Appendix E.

6.1. Diffusion equation with nonlinear reaction term

Let \( \Omega = (0, 1)^2 \) and \( \mathcal{P} = \mathbb{R}^2 \) and consider the PDE

\[
-\nabla^2 u(x_1, x_2; \theta) + g(u(x_1, x_2; \theta), \theta) = 100 \sin(2\pi x_1) \sin(2\pi x_2), \quad x \in \Omega, \tag{44}
\]

with homogeneous Dirichlet boundary conditions, where \( x = [x_1, x_2]^T, \theta = [\theta_1, \theta_2]^T \in \mathcal{P}, \) and \( u : \Omega \times \mathcal{P} \rightarrow \mathbb{R} \) is the solution function. The nonlinear reaction term \( g \) is

\[
g(u(x; \theta), \theta) = (0.1 \sin(\theta_1) + 2) \exp(-2.7\theta_2^2)(\exp(1.8\theta_2 u(x; \theta)) - 1). \]
The PDE (44) is discretized with finite differences on a grid with equidistant grid points and mesh width $h > 0$. The corresponding system of nonlinear equations is solved with Newton’s method and inexact line search based on the Armijo condition. The model $G_\ell : \mathcal{P} \to \mathcal{Y}$ derived with mesh width $h = 2^{-\ell-2}$ maps from $\mathcal{P}$ into $\mathcal{Y} = \mathbb{R}^{12}$. The components of the observed data $G_\ell(\theta) \in \mathcal{Y}$ correspond to the value of the approximated solution function at the spatial coordinates $[0.25i, 0.2j]^T \in \Omega$ with $i \in [3], j \in [4]$. We set $\theta^* = [-\pi/4, 3]^T$ and consider the data $y = G_{L+1}(\theta^*) + e$, where $L = 3$ (i.e., $h = 2^{-5}$) and $e$ adds zero-mean Gaussian noise of $0.5\%$. The prior distribution is a Gaussian distribution with mean $[\pi/2, 1.5]$ and diagonal covariance matrix with $[50, 0.5]$ on the diagonal.

**SVGD and MLSVGD** We start with $N = 1000$ particles sampled from a normal distribution with mean $[1, 1]^T$ and diagonal covariance matrix with $10^{-4}$ on the diagonal. The kernel is $k(\theta, \theta') = \exp(-||\theta - \theta'||^2/(2\sigma_k))$ with $\sigma_k = 10^{-2}$. The gradient of the likelihood is approximated with central differences with mesh width $2^{-6}$. The step size is $\delta = 10^{-1}$. We run SVGD for $\pi^{(L)}$ until the norm of the estimated gradient (43) reaches a tolerance $\epsilon$. We also run MLSVGD as in Algorithm 1 with levels $\ell \in \{1, 2, 3\}$ and $\ell \in \{1, 2, 3\}$.

**Results** Figure 2 shows the decay of the estimated gradient norm (43) for SVGD and MLSVGD with two and three levels, respectively, for a tolerance $\epsilon = 10^{-4}$. While the number of total iterations over all levels in MLSVGD is higher than in SVGD, the costs per iteration are lower on lower levels and thus MLSVGD achieves a speedup of about...
8 in this example. Notice that a switch to the next higher level leads to an increase of the gradient norm (e.g., Figure 2b near 1000 iterations), which is then reduced quickly in subsequent iterations. MLSVGD with 2 levels \((l \in \{1, 3\})\) achieves a slightly lower speedup than MLSVGD with 3 levels in this example. Figure 2b shows the speedup of MLSVGD with 3 levels for various tolerances. The speedup increases as the tolerance decreases. Figure 2c shows the error of the particle mean with respect to an MCMC reference over 10 replicates (cf. Appendix E). The proposed MLSVGD with 3 levels achieves more than one order of magnitude speedup compared to SVGD on the highest level. Notice that running SVGD on the lowest level MLSVGD with 3 levels achieves about the same speedup as MLSVGD with 3 levels in this example. Figure 3b-c show the pointwise error of the finite-difference solution \(u\) of (44) computed at the particle mean of MLSVGD and the particle mean of SVGD with the same costs as MLSVGD. The error is computed with respect to the solution at the MCMC reference. Notice the lighter color in the SVGD plot, which indicates higher pointwise error.

6.2. Euler-Bernoulli beam

Let \(\Omega = (0, 1) \subset \mathbb{R}\) and consider the Euler-Bernoulli beam described by

\[
\partial^2_x (E(x) \partial^2_x u(x)) = f(x), \quad x \in \Omega,
\]

where \(u : \Omega \to \mathbb{R}\) is the vertical deflection of the beam and \(f : \Omega \to \mathbb{R}\) is the load. The effective stiffness of the beam is given by \(E : \Omega \to \mathbb{R}\) and describes the beam geometry and material properties. The beam is in cantilever configuration, where the left boundary is fixed and the right boundary is free.

The observation \(y \in \mathbb{R}^{41}\) is the displacement at 41 equidistant points in \(\Omega\) polluted with 0.01% zero-mean Gaussian noise. We consider a smoothed piecewise constant approximation \(\tilde{E}_d\) of the stiffness \(E\) that depends on \(d \in \mathbb{N}\) parameters \(\theta = [\theta_1, \ldots, \theta_d]^T\), cf. Appendix E. The parameter-to-observable map \(G_d\) is then given by numerically solving (45) with stiffness \(\tilde{E}_d\) on level \(d\). The levels \(d = 1, \ldots, 6\) are corresponding to a discretization of the PDE on a mesh of 51, 101, \ldots, 501 equidistant grid points. The prior is log-normal with parameters \(\mu = 1\) and \(\sigma = 0.05\).

Results for SVGD and MLSVGD

The initial distribution is normal with mean \([1, 1, \ldots, 1]^T \in \mathbb{R}^d\) and diagonal covariance with \(4 \times 10^{-4}\) on the diagonal. We consider \(N = 500\) particles. The step size is \(\delta = 10^{-3}\) for \(d = 3\) and \(\delta = 10^{-2}\) for \(d \in \{6, 9\}\) and \(\delta = 5 \times 10^{-3}\) for \(d \in \{12, 16\}\). The kernel bandwidth \(\sigma_k\) is \(10^{-6}\) for \(d = 3\) and \(10^{-5}\) for \(d \in \{6, 9\}\) and \(5 \times 10^{-5}\) for \(d \in \{12, 16\}\). We consider MLSVGD for levels \(d \in \{1, \ldots, 6\}\) and \(d \in \{1, 3, 6\}\). The rest of the setup is the same as in Section 6.1. Figure 4 shows the convergence behavior of MLSVGD and SVGD for the problem with \(d = 9\) dimensions and tolerance \(\epsilon = 5 \times 10^{-3}\). A speedup of about 6 is observed to reach an estimated gradient norm below \(\epsilon\). Note that MLSVGD with 3 levels achieves about the same speedup as MLSVGD with 6 levels, which indicates that adding more and more intermediate levels cannot further reduce the costs. Speedups are reported in Figure 4b for MLSVGD with 3 levels; cf. Appendix E. If one asks for the error of the particle mean to be below \(3 \times 10^{-4}\) with respect to an MCMC reference, then MLSVGD achieves a speedup.
of about one order of magnitude compared to SVGD, as shown in Figure 4c. Figure 5 shows the relative pointwise error of the finite-difference solution $u$ of (45) computed at the particles obtained with MLSVGD and single-level SVGD; see also Figure 11 in the appendix. The error bars denote the minimum and maximum pointwise error of the inferred solutions over the ensemble of particles. The results show that MLSVGD achieves a similar error as single-level SVGD even though the computational costs of MLSVGD are lower than single-level SVGD in this example; cf. Figure 4b. Additionally, the variation of the error in terms of minimum and maximum error over the ensemble is comparable between MLSVGD and single-level SVGD in this example.

### 7. Conclusions

The proposed MLSVGD shows that speedups compared to single-level SVGD can be achieved by balancing the SVGD error with the discretization error given by a hierarchy of ever more accurate and ever more expensive-to-sample distributions. The analysis is conducted in the mean-field limit and shows a cost complexity reduction of MLSVGD compared to single-level SVGD. The numerical experiments demonstrate empirically that MLSVGD achieves up to one order of magnitude speedup compared to single-level SVGD in the discrete-time and finite-particle regime in the applications considered in this work. A cost analysis in discrete time and with finite particles remains future work for MLSVGD especially because there are only limited convergence results available even for single-level SVGD for discrete-time and finite-particle regimes.

### Acknowledgements

The first and third author were partially funded by the National Science Foundation under grants CMMI-1761068 and IIS-1901091. The first author was additionally supported in part by the Research Training Group in Modeling and Simulation funded by the National Science Foundation via grant RTG/DMS 1646339. The third author acknowledges additional support from the AFOSR MURI on multi-information sources of multi-physics systems under Award Number FA9550-15-1-0038 (Dr. Fariba Fahroo).

### References

[1] T. Alsup and B. Peherstorfer. Context-aware surrogate modeling for balancing approximation and sampling costs in multi-fidelity importance sampling and Bayesian inverse problems. \textit{arXiv:2010.11708}, 2020.

[2] D. Bakry, I. Gentil, and M. Ledoux. \textit{Analysis and Geometry of Markov Diffusion Operators}, volume 348 of \textit{Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]}. Springer, 2014.
[3] A. Beskos, A. Jasra, K. Law, R. Tempone, and Y. Zhou. Multilevel sequential Monte Carlo samplers. *Stochastic Processes and their Applications*, 127(5):1417 – 1440, 2017.

[4] W. Briggs, V. E. Henson, and S. McCormick. *A Multigrid Tutorial, Second Edition*. Society for Industrial and Applied Mathematics, second edition, 2000.

[5] H. Bungartz and M. Griebel. Sparse grids. *Acta Numerica*, 13:147–269, 2004.

[6] 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 H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 15130–15139. Curran Associates, Inc., 2019.

[7] S. Chewi, T. L. Gouic, C. Lu, T. Maunu, and P. Rigollet. SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020.

[8] J. A. Christen and C. Fox. Markov chain Monte Carlo using an approximation. *Journal of Computational and Graphical Statistics*, 14(4):795–810, 2005.

[9] K. A. Cliffe, M. Giles, R. Scheichl, and A. L. Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Computing and Visualization in Science*, 14(1):3–15, 2011.

[10] G. Detommaso, T. Cui, Y. Marzouk, A. Spantini, and R. Scheichl. A Stein variational Newton method. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 9169–9179. Curran Associates, Inc., 2018.

[11] T. J. Dodwell, C. Ketelsen, R. Scheichl, and A. L. Teckentrup. A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):1075–1108, 2015.

[12] A. Duncan, N. Nuesken, and L. Szpruch. On the geometry of Stein variational gradient descent, 2019.

[13] C. Fox and G. Nicholls. Sampling conductivity images via MCMC. In *The Art and Science of Bayesian Image Analysis*, pages 91–100. University of Leeds, 1997.

[14] M. Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

[15] A. Gregory, C. J. Cotter, and S. Reich. Multilevel ensemble transform particle filtering. *SIAM Journal on Scientific Computing*, 38(3):A1317–A1338, 2016.

[16] H. Haario, M. Laine, A. Mira, and E. Saksman. DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354, Dec. 2006.

[17] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 04 2001.

[18] W. Hackbush. *Multi-Grid Methods and Applications*. Springer, 1985.

[19] J. Han and Q. Liu. Stein variational adaptive importance sampling. *arXiv:1704.05201*, 2017.

[20] S. Heinrich. Multilevel Monte Carlo methods. In S. Margenov, J. Waśniewski, and P. Yalamov, editors, *Large-Scale Scientific Computing*, pages 58–67, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[21] V. H. Hoang, C. Schwab, and A. M. Stuart. Complexity analysis of accelerated MCMC methods for Bayesian inversion. *Inverse Problems*, 29(8):085010, jul 2013.

[22] H. Hoel, K. Law, and R. Tempone. Multilevel ensemble Kalman filtering. *SIAM Journal on Numerical Analysis*, 54(3):1813–1839, 2016.
[23] A. Jasra, K. Kamatani, K. Law, and Y. Zhou. Multilevel particle filters. *SIAM Journal on Numerical Analysis*, 55(6):3068–3096, 2017.

[24] J. Kaipio and E. Somersalo. Statistical inverse problems: Discretization, model reduction, and inverse crimes. *Journal of Computational and Applied Mathematics*, 198(2):493–504, 2007.

[25] A. Korba, A. Salim, M. Arbel, G. Luise, and A. Gretton. A non-asymptotic analysis for Stein variational gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[26] J. Latz, I. Papaioannou, and E. Ullmann. Multilevel sequential Monte Carlo for Bayesian inverse problems. *Journal of Computational Physics*, 368:154 – 178, 2018.

[27] C. Liu, J. Zhuo, P. Cheng, R. Zhang, and J. Zhu. Understanding and accelerating particle-based variational inference. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4082–4092, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[28] Q. Liu. Stein variational gradient descent as gradient flow. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 3115–3123. Curran Associates, Inc., 2017.

[29] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 2378–2386. Curran Associates, Inc., 2016.

[30] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012.

[31] Y. Marzouk and D. Xiu. A stochastic collocation approach to Bayesian inference in inverse problems. *Communications in Computational Physics*, 6(4):826–847, 2009.

[32] T. A. E. Moselhy and Y. M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815 – 7850, 2012.

[33] M. Parno and Y. Marzouk. Transport map accelerated Markov Chain Monte Carlo. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):645–682, 2018.

[34] B. Peherstorfer, B. Kramer, and K. Willcox. Multifidelity preconditioning of the cross-entropy method for rare event simulation and failure probability estimation. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):737–761, 2018.

[35] B. Peherstorfer and Y. Marzouk. A transport-based multifidelity preconditioner for Markov chain Monte Carlo. *Advances in Computational Mathematics*, 45:2321–2348, 2019.

[36] B. Peherstorfer, K. Willcox, and M. Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018.

[37] R. Ranganath, S. Gerrish, and D. Blei. Black Box Variational Inference. In S. Kaski and J. Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland, 22–25 Apr 2014.

[38] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 2015.
A. Metrics and divergences and other definitions

The Hellinger distance between two probability distributions $\mu$ and $\eta$ on $\mathbb{R}^d$ is defined as

$$d_{\text{Hell}}(\mu, \eta) = \sqrt{\frac{1}{2} \int_{\mathbb{R}^d} \left( \sqrt{\mu(\theta)} - \sqrt{\eta(\theta)} \right)^2 \, d\theta}.$$  (46)

The Kullback-Leibler (KL) divergence from $\mu$ to $\eta$ is defined as

$$\text{KL}(\mu \mid \mid \eta) = \int_{\mathbb{R}^d} \frac{\mu(\theta)}{\eta(\theta)} \log \left( \frac{\mu(\theta)}{\eta(\theta)} \right) \, d\theta.$$  (47)

Define the $L^2(\mu)$ space for a distribution $\mu$ and vector-valued functions as

$$L^2(\mu) = \left\{ f : \int_{\mathbb{R}^d} \| f(\theta) \|^2 \mu(\theta) \, d\theta < \infty \right\},$$  (48)

and the $L^2(\mu)$ norm of a vector-valued function as

$$\| f \|^2_{L^2(\mu)} = \int_{\mathbb{R}^d} \| f(\theta) \|^2 \mu(\theta) \, d\theta.$$  (49)

B. Lemma 1 and proof

**Lemma 1.** If Assumption 5 holds, there exists a constant $C > 0$ such that for all $1 \leq \ell_1, \ell_2 \leq \infty$ sufficiently large

$$\text{KL}(\pi^{(\ell_1)} \mid \mid \pi^{(\ell_2)}) \leq C \| G_{\ell_1} - G_{\ell_2} \|_{L^2(\pi_0)}.$$  (50)

*Note that for $\ell = \infty$ we say $G_{\ell} = G$.  

We note that this proof closely mirrors the proofs of Lemmas 4.2 and 4.3 in [31], but is slightly more general.

**Proof.** For brevity write $G_i = G_{\ell_i}$, $Z_i = Z_{\ell_i}$, and $\pi_i = \pi^{(\ell_i)}$ for $i = 1, 2$. Consider that for any vectors $u, v, w \in \mathbb{R}^d$ and symmetric positive definite matrix $A \in \mathbb{R}^{d \times d}$ we have

$$\| u - w \|_A^2 - \| v - w \|_A^2 = \| (u - v) + (v - w) \|_A^2 - \| v - w \|_A^2 = \langle (u - v), A(u - v) \rangle + 2 \langle (u - v), A(v - w) \rangle = \langle (u - v), A(u - v + 2w) \rangle \leq \| u - v \| \cdot \| A(u + v - 2w) \|,$$  (51)
with the last line following from the Cauchy-Schwarz inequality. Applying this bound with \( u = G_1(\theta) \), \( v = G_2(\theta) \), \( w = y \), and \( A = \Gamma^{-1} \) gives

\[
\int_\Theta \left| \left| y - G_1(\theta) \right| \right|^2_{\Gamma^{-1}} - \left| \left| y - G_2(\theta) \right| \right|^2_{\Gamma^{-1}} \pi_0(\theta) \, d\theta \\
\leq \int_\Theta \left| \left| G_1(\theta) - G_2(\theta) \right| \right| \cdot \left| \left| \Gamma^{-1}(2y - G_1(\theta) - G_2(\theta)) \right| \right| \pi_0(\theta) \, d\theta \\
\leq \left| \left| G_1 - G_2 \right| \right|_{L^2(\pi_0)} \cdot \left| \left| \Gamma^{-1}(2y - G_1 - G_2) \right| \right|_{L^2(\pi_0)},
\]

where the last line again follows from the Cauchy-Schwarz inequality on the inner-product space \( L^2(\pi_0) \). The KL divergence can now be bounded using Equation (52)

\[
\text{KL}(\pi_1 \mid \pi_2) = \int_\Theta \pi_1(\theta) \log \left( \frac{\pi_1(\theta)}{\pi_2(\theta)} \right) \, d\theta \\
= \int_\Theta \pi_1(\theta) \log \left( \frac{Z_2 \exp \left( -\frac{1}{2} \left| \left| y - G_1(\theta) \right| \right|^2_{\Gamma^{-1}} \right)}{Z_1 \exp \left( -\frac{1}{2} \left| \left| y - G_2(\theta) \right| \right|^2_{\Gamma^{-1}} \right)} \right) \, d\theta \\
= \log \left( \frac{Z_2}{Z_1} \right) + \int_\Theta \pi_1(\theta) \log \left( \frac{\exp \left( -\frac{1}{2} \left| \left| y - G_1(\theta) \right| \right|^2_{\Gamma^{-1}} \right)}{\exp \left( -\frac{1}{2} \left| \left| y - G_2(\theta) \right| \right|^2_{\Gamma^{-1}} \right)} \right) \, d\theta \\
\leq \log \left( \frac{Z_2}{Z_1} \right) + \frac{1}{2Z_1} \int_\Theta \left| \left| y - G_1(\theta) \right| \right|^2_{\Gamma^{-1}} - \left| \left| y - G_2(\theta) \right| \right|^2_{\Gamma^{-1}} \pi_0(\theta) \, d\theta \\
\leq \log \left( \frac{Z_2}{Z_1} \right) + \frac{1}{2Z_1} \left| \left| G_1 - G_2 \right| \right|_{L^2(\pi_0)} \cdot \left| \left| \Gamma^{-1}(2y - G_1 - G_2) \right| \right|_{L^2(\pi_0)},
\]

where in the second-to-last line we used the fact that \( \frac{1}{2} \left| \left| y - G_1(\theta) \right| \right|^2_{\Gamma^{-1}} \geq 0 \) and hence

\[
\exp \left( -\frac{1}{2} \left| \left| y - G_1(\theta) \right| \right|^2_{\Gamma^{-1}} \right) \leq 1.
\]

We bound the logarithm of the ratio of the normalizing constants by first bounding the difference of the normalizing constants using the bound in Equation (52)

\[
\left| Z_1 - Z_2 \right| = \left| \int_\Theta \left\{ \exp \left( -\frac{1}{2} \left| \left| y - G_1(\theta) \right| \right|^2_{\Gamma^{-1}} \right) - \exp \left( -\frac{1}{2} \left| \left| y - G_2(\theta) \right| \right|^2_{\Gamma^{-1}} \right) \right\} \pi_0(\theta) \, d\theta \\
\leq \int_\Theta \left| \exp \left( -\frac{1}{2} \left| \left| y - G_1(\theta) \right| \right|^2_{\Gamma^{-1}} \right) - \exp \left( -\frac{1}{2} \left| \left| y - G_2(\theta) \right| \right|^2_{\Gamma^{-1}} \right) \right| \pi_0(\theta) \, d\theta \\
\leq \frac{1}{2} \int_\Theta \left| \left| y - G_1(\theta) \right| \right|^2_{\Gamma^{-1}} - \left| \left| y - G_2(\theta) \right| \right|^2_{\Gamma^{-1}} \pi_0(\theta) \, d\theta \\
\leq \frac{1}{2} \left| \left| G_1 - G_2 \right| \right|_{L^2(\pi_0)} \cdot \left| \left| \Gamma^{-1}(2y - G_1 - G_2) \right| \right|_{L^2(\pi_0)}.
\]

The third line follows from the fact that \( |e^x - e^y| \leq |x - y| \) for all \( x, y \geq 0 \). Let \( \gamma_{\text{min}} > 0 \) denote the smallest eigenvalue of the noise covariance matrix \( \Gamma \). By the triangle inequality

\[
\left| \left| \Gamma^{-1}(2y - G_1 - G_2) \right| \right|_{L^2(\pi_0)} \leq 2 \left| \left| \Gamma^{-1}y \right| \right|_{L^2(\pi_0)} + \left| \left| \Gamma^{-1}(G_1 + G_2) \right| \right|_{L^2(\pi_0)} \\
\leq 2\left| \left| \Gamma^{-1}y \right| \right|_{L^2(\pi_0)} + 2\left| \left| \Gamma^{-1}G \right| \right|_{L^2(\pi_0)} + \left| \left| \Gamma^{-1}(G_1 + G_2 - 2G) \right| \right|_{L^2(\pi_0)} \\
\leq 2\left| \left| \Gamma^{-1}y \right| \right|_{L^2(\pi_0)} + 2\left| \left| \Gamma^{-1}G \right| \right|_{L^2(\pi_0)} + \frac{1}{\gamma_{\text{min}}} \left| \left| G_1 - G \right| \right|_{L^2(\pi_0)} + \frac{1}{\gamma_{\text{min}}} \left| \left| G_2 - G \right| \right|_{L^2(\pi_0)}.
\]

18
Since $\|G_\ell - G\|_{L^2(\pi_0)} \to 0$ by Assumption 5, we can bound $\|G_1 - G\|_{L^2(\pi_0)}$ and $\|G_2 - G\|_{L^2(\pi_0)}$ independently of $\ell_1$ and $\ell_2$. Therefore, there exists a constant $b_1 > 0$ independent of $\ell$ such that

$$\|\Gamma^{-1}(2y - G_1 - G_2)\|_{L^2(\pi_0)} \leq b_1.$$  \hfill (57)

Combining Equations (55) and (57) yields

$$|Z_1 - Z_2| \leq \frac{b_1}{2} \|G_1 - G_2\|_{L^2(\pi_0)}. \hfill (58)$$

The ratio of the normalizing constants can be written as

$$\left| \frac{Z_2}{Z_1} - 1 \right| = \frac{1}{Z_1} |Z_1 - Z_2|,$$

so the logarithm can be bounded as

$$\left| \log \left( \frac{Z_2}{Z_1} \right) \right| \leq \max \left\{ \left| \log \left( 1 - \frac{|Z_2 - Z_1|}{Z_1} \right) \right|, \left| \log \left( 1 + \frac{|Z_2 - Z_1|}{Z_1} \right) \right| \right\} \leq \frac{|Z_1 - Z_2|}{Z_1 - |Z_1 - Z_2|}, \hfill (60)$$

since $x \mapsto |\log x|$ is decreasing on $(0, 1]$ and increasing on $[1, \infty)$. Combining this with the inequality that $1 + x \leq \log(1 + x) \leq x$ for all $x > 0$ gives

$$\left| \log \left( \frac{Z_2}{Z_1} \right) \right| \leq \max \left\{ \frac{|Z_2 - Z_1|}{Z_1}, \frac{|Z_2 - Z_1|}{Z_1} \right\} \leq \frac{|Z_1 - Z_2|}{Z_1 - |Z_1 - Z_2|}.$$  \hfill (61)

Since $Z_\ell \to Z \in (0, \infty)$ is a convergent sequence, there exists a constant $b_2 > 0$ such that

$$Z_1^{-1} \sup_{\ell \geq 1} Z_\ell^{-1} \leq b_2. \hfill (62)$$

Moreover, for all $\ell_1, \ell_2$ sufficiently large $|Z_1 - Z_2| \leq b_2^{-1}/2$. Using the bound gives

$$\left| \log \left( \frac{Z_2}{Z_1} \right) \right| \leq \frac{|Z_1 - Z_2|}{b_2^{-1} - |Z_1 - Z_2|} \leq 2b_2 |Z_1 - Z_2|. \hfill (63)$$

Combining Equations (55), (57), (58), (62), and (63) gives

$$\text{KL}(\pi_1 || \pi_2) \leq \frac{3}{2} b_1 b_2 \|G_1 - G_2\|_{L^2(\pi_0)}. \hfill (64)$$

Now set $C = \frac{3}{2} b_1 b_2$ to obtain the result.

### C. Lemma 2 and proof

**Lemma 2.** If Assumption 5 holds with $\psi(\ell) = b_0 s^{-\alpha \ell}$, then Assumptions 2-4 also hold with the same rate $\alpha$.

**Proof.** Let $\ell_1 = \ell$ and $\ell_2 = \infty$, so that by Lemma 1 in Appendix B we immediately have

$$\text{KL}(\pi(\ell) || \pi) \leq C \|G_\ell - G\|_{L^2(\pi_0)} \leq C \psi(\ell) = C b_0 s^{-\alpha \ell}, \hfill (65)$$

so that $k_1 = C b_0$. Moreover, setting $\ell_1 = \ell - 1$ and $\ell_2 = \ell$ and using the triangle inequality gives

$$\text{KL}(\pi(\ell - 1) || \pi(\ell)) \leq C \|G_{\ell-1} - G_\ell\|_{L^2(\pi_0)} \leq C \left( \|G_{\ell-1} - G_\ell\|_{L^2(\pi_0)} + \|G_\ell - G\|_{L^2(\pi_0)} \right). \hfill (66)$$

Thus,

$$\text{KL}(\pi(\ell - 1) || \pi(\ell)) \leq C \left( 1 + \frac{\psi(\ell - 1)}{\psi(\ell)} \right) \psi(\ell) \leq C b_0 (1 + s^\alpha) s^{-\alpha \ell}, \hfill (67)$$

so that $k_2 = C b_0 (1 + s^\alpha)$. \hfill \(\Box\)
D. A triangle-like inequality for the KL divergence

Let $\rho_0, \rho_1, \rho_2$ be three probability distributions on $\Theta$. We have that

$$KL(\rho_0 \mid \mid \rho_2) = \int_{\Theta} \rho_0(\theta) \log \left( \frac{\rho_0(\theta)}{\rho_2(\theta)} \right) d\theta$$

$$\quad = \int_{\Theta} \rho_0(\theta) \log \left( \frac{\rho_0(\theta)}{\rho_1(\theta)} \frac{\rho_1(\theta)}{\rho_2(\theta)} \right) d\theta$$

$$\quad = \int_{\Theta} \rho_0(\theta) \log \left( \frac{\rho_0(\theta)}{\rho_1(\theta)} \right) d\theta + \int_{\Theta} \rho_0(\theta) \log \left( \frac{\rho_1(\theta)}{\rho_2(\theta)} \right) d\theta$$

$$\quad = KL(\rho_0 \mid \mid \rho_1) + KL(\rho_1 \mid \mid \rho_2) + \int_{\Theta} (\rho_0(\theta) - \rho_1(\theta)) \log \left( \frac{\rho_1(\theta)}{\rho_2(\theta)} \right) d\theta,$$

(cf. the inequality given in [31]). We refer to this third term in the last line as the remainder term.

E. Details about numerical experiments

E.1. General

The step size $\delta$ and kernel bandwidth $h$ was chosen via a manual process so that SVGD on the highest level numerically converged. The same $\delta$ and $h$ are used for SVGD and MLSVGD. Time measurements were performed on compute nodes with Intel Xeon CPU E5-2690 v2, restricted to 8 cores and 32GB memory, with a Matlab implementation. The MCMC reference is computed with the delayed-rejection adaptive Metropolis (DRAM) method [17,16] on the highest level $L$ of the respective problem. The covariance matrix of the Gaussian proposal is initialized to be diagonal with $10^{-2}$ on the diagonal. The burn-in time is 10,000 samples. Another 20,000 samples are generated and every other sample is then used to compute the MCMC reference mean $\bar{\theta}$ of the parameter. The error reported in Figure 3a and Figure 4c is $\frac{1}{10} \sum_{i=1}^{10} \| \bar{\theta} - \theta^{(i)} \|_2$, where $\theta^{(i)}$ is the mean of (ML)SVGD particles of the $i$-th replicate.

E.2. Diffusion equation with nonlinear reaction term

We repeat the experiments of Section 6.1 with $N \in \{500, 2500, 5000\}$ particles and show the corresponding speedups in Figure 6. The speedup of MLSVGD is roughly the same over the different numbers of particles, which is expected because the cost of MLSVGD scales with the number of particles as the cost of SVGD.
E.3. Euler-Bernoulli beam

In Section 6.2 we consider the PDE (45) for \( \Omega = [0, 1] \), where \( u : \Omega \to \mathbb{R} \) is the vertical deflection of the beam and \( f : \Omega \to \mathbb{R} \) is the load. The effective stiffness of the beam is given by \( E : \Omega \to \mathbb{R} \) and describes beam geometry and material properties. The beam is in cantilever configuration, where the left boundary is fixed and the right boundary is free i.e., the boundary conditions are

\[
    u(0) = 0, \quad \frac{\partial u}{\partial x} \big|_{x=0} = 0, \quad \frac{\partial^3 u}{\partial x^3} \big|_{x=1} = 0, \quad \frac{\partial^3 u}{\partial x^3} \big|_{x=1} = 0.
\]

We use the same stiffness \( E \) available in the model developed by Matthew Parno for the 2018 Gene Golub SIAM Summer School on “Inverse Problems: Systematic Integration of Data with Models under Uncertainty.” The model is available on GitHub.\(^1\)

**Forward model**

The forward model is derived as follows. Consider the function \( I : \mathbb{R} \times \Omega \to \mathbb{R} \) defined as

\[
    I(x, \alpha) = \left( 1 + \exp \left( \frac{-x - \alpha}{0.005} \right) \right)^{-1},
\]

with

\[
\lim_{x \to -\infty} I(x, \alpha) = 0, \quad \lim_{x \to +\infty} I(x, \alpha) = 1
\]

such that there is a smooth transition from 0 to 1 at \( \alpha \). For \( k > 1 \), let \( \alpha_1, \ldots, \alpha_{k+1} \) be \( k+1 \) equidistant points in \( \Omega \). Let \( \mathbb{R}_+ = \{ z \in \mathbb{R} : z > 0 \} \) and consider the parameter \( \theta = [\theta_1, \ldots, \theta_k]^T \in \mathbb{R}_+^k \). Define the function \( \hat{E}_k : \Omega \times \mathbb{R} \to \mathbb{R} \) as

\[
\hat{E}_k(x, \theta) = (1 - I(x, \alpha_1)) \hat{E}_k(x, \theta) = (1 - I(x, \alpha_1)) \hat{E}_k(x, \theta_{i-1})
\]

Given a parameter \( \theta \), the function \( \hat{E}_k(x) \) is a smooth approximation of the piecewise constant function \( \sum_{i=1}^{k} \theta_i \mathbb{I}(\alpha_i, \alpha_{i+1}] \), where \( \mathbb{I}(\alpha_i, \alpha_{i+1}] \) is the indicator function of the interval \( (\alpha_i, \alpha_{i+1}] \subset \mathbb{R} \).

**Additional plots for \( d \in \{3, 6, 12, 16\} \)**

Figure 7–10 show the analogous results to Figure 4 for dimension \( d \in \{3, 6, 12, 16\} \), respectively. Figure 11 shows the analogous results to Figure 5. The behavior of MLSVGD compared to SVGD is qualitatively the same as for dimension \( d = 9 \).

---

\(^1\) https://github.com/g2s3-2018/labs
Figure 8.: Euler-Bernoulli beam: Results of MLSVGD for dimension $d = 6$.

Figure 9.: Euler-Bernoulli beam: Results of MLSVGD for dimension $d = 12$.

Figure 10.: Euler-Bernoulli beam: Results of MLSVGD for dimension $d = 16$. 

22
Figure 11.: Euler-Bernoulli beam: Minimum and maximum of pointwise error over ensemble of inferred solutions for $d \in \{3, 6, 12, 16\}$. 

(a) $d = 3$, single-level SVGD
(b) $d = 3$, MLSVGD (3 levels)
(c) $d = 3$, MLSVGD (6 levels)
(d) $d = 6$, single-level SVGD
(e) $d = 6$, MLSVGD (3 levels)
(f) $d = 6$, MLSVGD (6 levels)
(g) $d = 12$, single-level SVGD
(h) $d = 12$, MLSVGD (3 levels)
(i) $d = 12$, MLSVGD (6 levels)
(j) $d = 16$, single-level SVGD
(k) $d = 16$, MLSVGD (3 levels)
(l) $d = 16$, MLSVGD (6 levels)