Proceedings of Machine Learning Research vol 145:1–25, 2021 2nd Annual Conference on Mathematical and Scientific Machine Learning

Multilevel Stein variational gradient descent with applications to Bayesian inverse problems

Terrence Alsup ALSUP@CIMS.NYU.EDU
Luca Venturi VENTURI@CIMS.NYU.EDU
Benjamin Peherstorfer PEHERSTO@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences, New York University

Editors: Joan Bruna, Jan S Hesthaven, Lenka Zdeborova

Abstract

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 π 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 (Robert and Casella, 2004; Ranganath et al., 2014; Rezende and Mohamed, 2015; Zhang et al., 2019). 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 π for increasing level $\ell \to \infty$. Here, intractable means that one cannot numerically evaluate the (unnormalized) density of π. For example, one finds such a setup in Bayesian inverse problems (Stuart, 2010; Kaipio and Somersalo, 2007; Martin et al., 2012), where the target π 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 π. To approximately sample from the target π, one then selects a level L such that $\pi^{(L)}$ is a sufficiently accurate approximation of π and then applies Monte Carlo or particle methods to $\pi^{(L)}$; see, e.g., (Stuart, 2010; Kaipio and Somersalo, 2007; Martin et al., 2012). 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.

Our contributions We propose to extend Stein variational gradient descent (SVGD) (Liu and Wang, 2016) to a multilevel SVGD (MLSVGD) that leverages the distributions from all levels $\ell = 1, \dots, L$ to more efficiently approximately sample from π 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 $\ell = 1, \dots, L$ 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 (Hackbush, 1985; Briggs et al., 2000), sparse grid approximations (Bungartz and Griebel, 2004), multilevel Monte Carlo for estimating statistics (Heinrich, 2001; Giles, 2008; Cliffe et al., 2011); and multifidelity methods that leverage low-fidelity models without clear hierarchies (Peherstorfer et al., 2018b). 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 (Christen and Fox, 2005; Fox and Nicholls, 1997), multilevel Metropolis–Hastings (Dodwell et al., 2015); and MCMC methods with importance sampling (Hoang et al., 2013). Then, there are multilevel/multifidelity variational methods, where a transport map (flow) is parametrized a priori; for example, (Alsup and Peherstorfer, 2020; Peherstorfer and Marzouk, 2019) build on (Moselhy and Marzouk, 2012; Parno and Marzouk, 2018) 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 (Jasra et al., 2017) and multilevel sequential Monte Carlo (Beskos et al., 2017) methods, ensemble Kalman filtering (Hoel et al., 2016), and extensions to nonlinear filtering using transport (Gregory et al., 2016); 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 (Latz et al., 2018; Wagner et al., 2020) and the multilevel cross-entropy method (Peherstorfer et al., 2018a) 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 Liu and Wang (2016) and further theoretically analyzed in (Liu, 2017); extended to consider Newton directions (Detommaso et al., 2018); exploiting geometry (Chen et al., 2019), and other acceleration techniques (Liu et al., 2019). 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 (Liu, 2017; Duncan et al., 2019) shows the mean-field limit. The work (Korba et al., 2020) shows non-asymptotic results. Further, the work (Chewi et al., 2020) 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 ℓ to divergence of the corresponding posterior distributions $\pi^{(\ell)}$ with respect to the intractable target π, where we build on results by Stuart (2010) and an inequality involving the Kullback-Leibler (KL) divergence introduced by Marzouk and Xiu (2009).

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 σ-algebra of Θ. Consider approximating some target measure η on Θ via an empirical measure, i.e., an ensemble of samples (particles); in the following, the distribution η, and all other distributions that will be considered, admit a density with respect to the Lebesgue measure over Θ. Moreover, the target distribution has the form $\eta \propto e^{-V}$ with the potential V.

2.1. Approximating measures with SVGD

The SVGD method (Liu and Wang, 2016) iteratively moves forward an empirical distribution given by an ensemble $\{\boldsymbol{\theta}_t^{[i]}\}_{i=1}^N$ from time t to time $t+\delta$ via a map $\boldsymbol{\phi}_t(\boldsymbol{\theta}) = \boldsymbol{\theta} - \delta \boldsymbol{g}_t(\boldsymbol{\theta})$, where δ is a step size and $\boldsymbol{g}_t : \Theta \to \mathbb{R}^d$ is a vector field. SVGD chooses $\boldsymbol{g}_t$ 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 μ_t and define the functional $J_t(\boldsymbol{g}) = \mathrm{KL}((I - \boldsymbol{g})_{\#}\mu_t || \eta)$, where $(I - \boldsymbol{g})_{\#}\mu_t$ denotes the pushfoward measure. Then, SVGD chooses the gradient by setting $\boldsymbol{g}_t = \nabla J_t(\mathbf{0})$, where $\mathbf{0}$ is the zero function. Using the RKHS formulation, there is a closed form expression for $\nabla J_t(\mathbf{0})$, so that during the gradient descent the particles evolve according to the ordinary differential equation (ODE)

$$\dot{\boldsymbol{\theta}}_t^{[i]} = -\nabla J_t(\mathbf{0}) \left(\boldsymbol{\theta}_t^{[i]}\right) = \mathbb{E}_{\boldsymbol{\theta}' \sim \mu_t} \left[K(\boldsymbol{\theta}', \boldsymbol{\theta}_t^{[i]}) \nabla \log \eta(\boldsymbol{\theta}') + \nabla_1 K(\boldsymbol{\theta}', \boldsymbol{\theta}_t^{[i]})\right] , \tag{1}$$

where ∇_1 denotes the gradient with respect to the first argument. In practice, the expectation is approximated using the empirical distribution of the ensemble of particles $\{\boldsymbol{\theta}_t^{[i]}\}_{i=1}^N$ and the ODE is integrated using the forward Euler method. Thus, the SVGD update becomes

$$\boldsymbol{\theta}_{t+\delta}^{[i]} = \boldsymbol{\theta}_t^{[i]} + \frac{\delta}{N} \left(\sum\nolimits_{j=1}^N \nabla_1 K(\boldsymbol{\theta}_t^{[j]}, \boldsymbol{\theta}_t^{[i]}) + \sum\nolimits_{j=1}^N K(\boldsymbol{\theta}_t^{[j]}, \boldsymbol{\theta}_t^{[i]}) \nabla \log \eta(\boldsymbol{\theta}_t^{[j]}) \right) . \tag{2}$$

In (Liu, 2017), the distribution of the particles $\{\boldsymbol{\theta}_t^{[i]}\}_{i=1}^N$ in the limit as $N \to \infty$ is given by the mean-field PDE

$$\partial_t \mu_t(\boldsymbol{\theta}) = -\nabla \cdot \left(\mu_t(\boldsymbol{\theta}) \mathbb{E}_{\boldsymbol{\theta}' \sim \mu_t} \left[K(\boldsymbol{\theta}', \boldsymbol{\theta}) \nabla \log \eta(\boldsymbol{\theta}') + \nabla_1 K(\boldsymbol{\theta}', \boldsymbol{\theta})\right]\right) , \tag{3}$$

with an initial measure μ_0; see also (Chewi et al., 2020; Han and Liu, 2017). Liu (2017) shows that a steady state is reached in the limit $t \to \infty$ and the empirical distribution converges weakly (i.e. in distribution) to the target η.

2.2. Approximating intractable target measures with SVGD

Consider now an intractable target distribution π; in contrast to the measure η in Section 2.1, we can neither evaluate the (unnormalized) density of π nor sample from π directly. Thus, the SVGD algorithm cannot directly be applied to π. Instead, suppose we have a sequence of distributions $(\pi^{(\ell)})_{\ell \geq 1}$ that converges weakly to π for $\ell \to \infty$ (note that weak convergence is implied by convergence in the KL divergence) and call ℓ the level. Moreover, we can evaluate the unnormalized density of each $\pi^{(\ell)}$ with computational costs c_ℓ. Such a setup is common in Bayesian inverse problems; cf. Section 1.

Single-level approximation with SVGD The aim is deriving a distribution μ that approximates π with accuracy ϵ. To quantify how close the approximation μ is to the target distribution, we consider the Hellinger distance $\mathrm{d}_{\mathrm{Hell}}(\cdot, \cdot)$ in the following: First, select a level $L \in \mathbb{N}$ such that $\mathrm{d}_{\mathrm{Hell}}(\pi^{(L)}, \pi) \leq \epsilon/2$. Then, to approximate $\pi^{(L)}$ with SVGD, derive μ with accuracy $\mathrm{d}_{\mathrm{Hell}}(\mu, \pi^{(L)}) \leq \epsilon/2$ from an initial distribution μ_0; the triangle inequality leads to $\mathrm{d}_{\mathrm{Hell}}(\mu, \pi) \leq \epsilon$. 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_L of evaluating the density $\pi^{(L)}$ on level L, which is independent of SVGD, and (2) the costs of SVGD to find μ from $\pi^{(L)}$ with initial distribution μ_0 to achieve $\mathrm{d}_{\mathrm{Hell}}(\mu, \pi^{(L)}) \leq \epsilon/2$. In the continuous SVGD given by Equation (1), we identify the costs of the approximation μ_T after integrating up to end time T as

$$c_{\mathrm{SL}}(T) = c_L T\,. \tag{4}$$

We will see that the integration time T depends on the divergence between μ_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 $\mathrm{d}_{\mathrm{Hell}}$ 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*

$$2\,\mathrm{d}_{\mathrm{Hell}}(\rho_1, \rho_2)^2 \leq \mathrm{KL}(\rho_1 \,||\, \rho_2) \tag{5}$$

for two distributions ρ_1, ρ_2; see Lemma 2.4 of (Tsybakov, 2009) (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 (Stuart, 2010).

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

We propose MLSVGD that leverages the measures $\pi^{(1)}, \dots, \pi^{(L-1)}$ 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.

single-level SVGD:

$$\begin{tikzcd}
\mu_0 \arrow[rrrrrrrr, "\pi^{(L)}", "T"'] & & & & & & & & \mu^{\mathrm{SL}}
\end{tikzcd}$$

proposed MLSVGD:

$$\begin{tikzcd}
\mu_0 \arrow[r, "\pi^{(1)}", "T_1"'] & \mu_{T_1}^{(1)} \arrow[r, "\pi^{(2)}", "T_2"'] & \mu_{T_2}^{(2)} \arrow[r, "\pi^{(3)}", "T_3"'] & \cdots \arrow[r, "\pi^{(L)}", "T_L"'] & \mu^{\mathrm{ML}}
\end{tikzcd}$$

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.

3.1. Continuous MLSVGD

To describe the proposed MLSVGD, consider the levels $\ell = 1, \dots, L$ and let μ_0 be an initial distribution. At level $\ell = 1$, we define $\mu_{T_1}^{(1)}$ as the distribution of the continuous SVGD (3) at time T_1 with target $\pi^{(1)}$ and initial μ_0. At level $\ell = 2$, we obtain $\mu_{T_2}^{(2)}$ at time T_2 with the target $\pi^{(2)}$ and initial distribution $\mu_{T_1}^{(1)}$. In general, at level ℓ, we obtain $\mu_{T_\ell}^{(\ell)}$ at time T_ℓ with target $\pi^{(\ell)}$ and initial distribution $\mu_{T_{\ell-1}}^{(\ell-1)}$. Thus, deriving $\mu_{T_\ell}^{(\ell)}$ is an iterative process over the levels $1, \dots, \ell - 1$, depicted in Figure 3, of first computing $\mu_{T_1}^{(1)}, \dots, \mu_{T_{\ell-1}}^{(\ell-1)}$. The costs of MLSVGD are given by

$$c_{\mathrm{ML}}(T_1, \dots, T_L) = \sum_{\ell=1}^{L} c_\ell T_\ell \,, \tag{6}$$

cf. the costs $c_{\mathrm{SL}}(T) = c_L T$ 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 $\mathrm{d}_{\mathrm{Hell}}(\mu, \pi) \leq \epsilon$.

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

$$c_\ell \leq c_0 s^{\gamma \ell} \,, \qquad \ell \in \mathbb{N} \,,$$

with constants $c_0, \gamma > 0$ independent of ℓ and $s > 1$.

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

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

where s is the same constant independent of ℓ as in Assumption 1 and μ_0 is the initial distribution.

Assumption 3 *There exists a rate $\lambda > 0$ such that for any initial distribution ν_0*

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

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

Korba et al. (2020) show that Assumption 3 is satisfied if the measures $\pi^{(\ell)}$ satisfy a Stein log-Sobolev inequality. Chewi et al. (2020) 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 (Bakry et al., 2014, Theorem 5.2.1); however, SVGD approximates the gradient in an RKHS and thus (Bakry et al., 2014, 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 $\mathrm{d}_{\mathrm{Hell}}(\pi^{(L)}, \pi) \leq \epsilon/2$ and then starts with a μ_0 to find μ_T^{SL} that satisfies $\mathrm{d}_{\mathrm{Hell}}(\mu_T^{\mathrm{SL}}, \pi^{(L)}) \leq \epsilon/2$. For brevity, we write $\mu^{\mathrm{SL}} = \mu_T^{\mathrm{SL}}$. The following proposition bounds the costs of this single-level SVGD with respect to the tolerance ϵ.

Proposition 2 *If Assumptions 1–3 hold, then the costs of continuous single-level SVGD to obtain μ^{SL} with*

$$\mathrm{d}_{\mathrm{Hell}}(\mu^{\mathrm{SL}}, \pi) \leq \epsilon$$

is bounded as

$$c_{\mathrm{SL}}^*(\epsilon) \leq \frac{2c_0 s^\gamma}{\lambda}\left(\frac{\sqrt{2k_1}}{\epsilon}\right)^{2\gamma/\alpha} \log\left(\frac{\sqrt{\mathrm{KL}(\mu_0 \,||\, \pi^{(L)})}}{\sqrt{2}\epsilon}\right)\,. \tag{7}$$

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

$$\mathrm{d}_{\mathrm{Hell}}(\mu^{\mathrm{SL}}, \pi) \leq \mathrm{d}_{\mathrm{Hell}}(\mu^{\mathrm{SL}}, \pi^{(L)}) + \mathrm{d}_{\mathrm{Hell}}(\pi^{(L)}, \pi),$$

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

$$\mathrm{d}_{\mathrm{Hell}}(\mu^{\mathrm{SL}}, \pi^{(L)}) \leq \sqrt{\frac{\mathrm{KL}(\mu^{\mathrm{SL}} \,||\, \pi^{(L)})}{2}}, \tag{8}$$

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

$$L = \left\lceil \frac{1}{\alpha}\log_s\left(\frac{2k_1}{\epsilon^2}\right)\right\rceil \leq \frac{1}{\alpha}\log_s\left(\frac{2k_1}{\epsilon^2}\right) + 1, \tag{9}$$

so that

$$\mathrm{d}_{\mathrm{Hell}}(\pi^{(L)}, \pi) \leq \sqrt{\frac{\mathrm{KL}(\pi^{(L)} \,||\, \pi)}{2}} \leq \sqrt{\frac{k_1 s^{-\alpha L}}{2}} \leq \frac{\epsilon}{2}. \tag{10}$$

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

$$T_{\mathrm{SL}}^* \leq \frac{1}{\lambda}\log\left(\frac{\mathrm{KL}(\mu_0 \,||\, \pi^{(L)})}{2\epsilon^2}\right). \tag{11}$$

The total cost to integrate until time T^*_{SL} at level L is thus

$$c^*_{\mathrm{SL}}(\epsilon) = c_0 s^{\gamma L} T^*_{\mathrm{SL}} \leq \frac{2c_0 s^\gamma}{\lambda} \left(\frac{\sqrt{2k_1}}{\epsilon}\right)^{2\gamma/\alpha} \log\left(\frac{\sqrt{\mathrm{KL}(\mu_0 \,||\, \pi^{(L)})}}{\sqrt{2}\epsilon}\right) .$$

■

Discussion of cost complexity of single-level SVGD The bound (7) in Proposition 2 shows that if we start with an initial distribution μ_0 that has a large KL divergence $\mathrm{KL}(\mu_0||\pi^{(L)})$ 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 μ_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 3.

Assumption 4 *There exists a constant $k_2 > 0$ independent of ℓ such that $\mathrm{KL}(\pi^{(\ell-1)}||\pi^{(\ell)}) \leq k_2 s^{-\alpha\ell}$, where α 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,

$$\mathrm{KL}\left(\mu^{(\ell-1)}_{T_{\ell-1}} \,||\, \pi^{(\ell)}\right) = \mathrm{KL}\left(\mu^{(\ell-1)}_{T_{\ell-1}} \,||\, \pi^{(\ell-1)}\right) + \mathrm{KL}\left(\pi^{(\ell-1)} \,||\, \pi^{(\ell)}\right) + R_\ell \tag{12}$$

with the remainder R_ℓ given by

$$R_\ell = \int_{\mathbb{R}^d} \left(\mu^{(\ell-1)}_{T_{\ell-1}}(\boldsymbol{\theta}) - \pi^{(\ell-1)}(\boldsymbol{\theta})\right) \log\left(\frac{\pi^{(\ell-1)}(\boldsymbol{\theta})}{\pi^{(\ell)}(\boldsymbol{\theta})}\right) \mathrm{d}\boldsymbol{\theta}. \tag{13}$$

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

Proposition 3 *If Assumptions 1–4 hold, then continuous MLSVGD gives μ^{ML} with $\mathrm{d}_{\mathrm{Hell}}(\mu^{ML}, \pi) \leq \epsilon$ with costs bounded as*

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

where R bounds (13).

Proof As in Equation (9) in the proof of Proposition 2 we select the level L as

$$L = \left\lceil \frac{1}{\alpha} \log_s \left(\frac{2k_1}{\epsilon^2} \right) \right\rceil \leq \frac{1}{\alpha} \log_s \left(\frac{2k_1}{\epsilon^2} \right) + 1, \tag{14}$$

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

$$c^*_{\mathrm{ML}}(\epsilon) = \sum_{\ell=1}^{L} c_0 s^{\gamma \ell} T_\ell \,, \tag{15}$$

where it remains to choose the integration times T_ℓ 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

$$\mathrm{KL}(\mu^{(\ell)}_{T_\ell} \,||\, \pi^{(\ell)}) \leq e^{-\lambda T_\ell} \left(\mathrm{KL} \left(\mu^{(\ell-1)}_{T_{\ell-1}} \,||\, \pi^{(\ell-1)} \right) + \mathrm{KL} \left(\pi^{(\ell-1)} \,||\, \pi^{(\ell)} \right) + R_\ell \right) \,, \tag{16}$$

giving a recursive bound on the KL divergence in terms of the KL divergence at the previous level. At each level ℓ choose the integration time T_ℓ so that

$$\mathrm{KL} \left(\mu^{(\ell)}_{T_\ell} \,||\, \pi^{(\ell)} \right) \leq \frac{\epsilon^2}{2} \tag{17}$$

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

$$\mathrm{KL} \left(\mu^{(\ell)}_{T_\ell} \,||\, \pi^{(\ell)} \right) \leq e^{-\lambda T_\ell} \left(\frac{\epsilon^2}{2} + \mathrm{KL}(\pi^{(\ell-1)} \,||\, \pi^{(\ell)}) + R_\ell \right) \leq \frac{\epsilon^2}{2} \,. \tag{18}$$

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

$$T_\ell \leq \frac{1}{\lambda} \log \left(1 + \frac{2(\mathrm{KL}(\pi^{(\ell-1)} \,||\, \pi^{(\ell)}) + R_\ell)}{\epsilon^2} \right) \,. \tag{19}$$

Finally, the total cost can be bounded by

$$c^*_{\mathrm{ML}}(\epsilon) \leq \sum_{\ell=1}^{L} \frac{c_0}{\lambda} s^{\gamma \ell} \log \left(1 + \frac{2(\mathrm{KL}(\pi^{(\ell-1)} \,||\, \pi^{(\ell)}) + R_\ell)}{\epsilon^2} \right) \,. \tag{20}$$

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

$$c^*_{\mathrm{ML}}(\epsilon) \leq \sum_{\ell=1}^{L} \frac{c_0}{\lambda} s^{\gamma \ell} \log \left(1 + \frac{2(k_2 + R)}{\epsilon^2} \right) \,. \tag{21}$$

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

$$\begin{aligned} c^*_{\mathrm{ML}}(\epsilon) &\leq \int_0^{L+1} \frac{c_0}{\lambda} s^{\gamma x} \log\left(1 + \frac{2(k_2+R)}{\epsilon^2}\right)\, \mathrm{d}x \\ &= \frac{c_0}{\lambda\gamma\log(s)} \log\left(1 + \frac{2(k_2+R)}{\epsilon^2}\right) s^{\gamma(L+1)} \\ &\leq \frac{2c_0 s^{2\gamma}}{\lambda\gamma\log(s)} \left(\frac{\sqrt{2k_1}}{\epsilon}\right)^{2\gamma/\alpha} \log\left(\frac{\sqrt{\epsilon^2 + 2(k_2+R)}}{\epsilon}\right). \end{aligned} \tag{22}$$

■

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_ℓ at each level. In particular, it allows us to set T_ℓ such that

$$\mathrm{KL}(\mu^{(\ell)}_{T_\ell} \,||\, \pi^{(\ell)}) \sim s^{-\alpha\ell}\,, \tag{23}$$

which leads to the following proposition that shows an improved cost complexity compared to Proposition 3.

Proposition 4 *If Assumptions 1–4 hold and $R_\ell \leq k_3 s^{-\alpha\ell}$, then the costs of continuous MLSVGD to have $\mathrm{d}_{\mathrm{Hell}}(\mu^{\mathrm{ML}},\, \pi) \leq \epsilon$ can be bounded as*

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

Proof Starting from Equation (16) in the proof of Proposition 3 change ϵ to instead be

$$\epsilon_\ell = \sqrt{2k_1} s^{-\alpha\ell/2} \tag{25}$$

at each level ℓ. By Assumption 2 we know that L is chosen so that

$$\epsilon_L^2 = 2k_1 s^{-\alpha L} \leq \epsilon^2\,, \tag{26}$$

so that $\epsilon_L \leq \epsilon$. Plugging in this choice gives that the integration times needed are

$$T^*_\ell \leq \frac{1}{\lambda}\log\left(s^\alpha + \frac{\mathrm{KL}(\pi^{(\ell-1)} \,||\, \pi^{(\ell)}) + R_\ell}{k_1 s^{-\alpha\ell}}\right). \tag{27}$$

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

$$T^*_\ell \leq \frac{1}{\lambda}\log\left(s^\alpha + \frac{k_2+k_3}{k_1}\right), \tag{28}$$

so that the integration time is fixed at each level. The cost is now bounded by

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

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

$$c^*_{\mathrm{ML}}(\epsilon) \leq \sum_{\ell=1}^{L} \frac{c_0}{\lambda} s^{\gamma \ell} \log\left(s^\alpha + \frac{k_2 + k_3}{k_1}\right) \leq \int_1^{L+1} \frac{c_0}{\lambda} s^{\gamma x} \log\left(s^\alpha + \frac{k_2 + k_3}{k_1}\right) \mathrm{d}x \,. \tag{30}$$

Computing the integral gives

$$c^*_{\mathrm{ML}}(\epsilon) \leq \frac{c_0 s^{2\gamma}}{\lambda \gamma \log(s)} \log\left(s^\alpha + \frac{k_2 + k_3}{k_1}\right) s^{\gamma L} \,. \tag{31}$$

Finally, by plugging in L we obtain

$$c^*_{\mathrm{ML}}(\epsilon) \leq \frac{c_0 s^{2\gamma}}{\lambda \gamma \log(s)} \log\left(s^\alpha + \frac{k_2 + k_3}{k_1}\right) \left(\frac{\sqrt{2k_1}}{\epsilon}\right)^{2\gamma/\alpha} \,. \tag{32}$$

■

Discussion of cost complexity of MLSVGD Looking at the single-level SVGD and MLSVGD cost bounds from Propositions 2 and 4, respectively, we note two major differences. The first is that there is no $\log \epsilon^{-1}$ term in the cost bound (24) of Proposition 4 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 ℓ 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 μ_0 and the target π, 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 5 *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 $\boldsymbol{\theta}$ from some noisy observed data $\mathbf{y} = G(\boldsymbol{\theta}^*) + \boldsymbol{e}$ with G denoting the parameter-to-observable map and $\boldsymbol{e}$ being the noise; see, e.g., (Stuart, 2010; Kaipio and Somersalo, 2007; Martin et al., 2012). Let π_0 be the prior and consider zero-mean Gaussian noise with covariance $\boldsymbol{\Gamma}$, then the posterior is given by

$$\pi(\boldsymbol{\theta}) = \frac{1}{Z} \exp\left(-\frac{1}{2}\|\mathbf{y} - G(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right) \pi_0(\boldsymbol{\theta}) \,, \tag{33}$$

with the normalizing constant

$$Z = \int_{\Theta} \exp\left(-\frac{1}{2}\|\mathbf{y} - G(\boldsymbol{\theta})\|^2_{\mathbf{\Gamma}^{-1}}\right) \pi_0(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}\,, \tag{34}$$

where $\|\boldsymbol{u}\|_{\mathbf{\Gamma}^{-1}} = \langle \mathbf{\Gamma}^{-1}\boldsymbol{u}, \boldsymbol{u}\rangle$. 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_ℓ and define Z_ℓ 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*

$$\|G(\boldsymbol{\theta}) - G_\ell(\boldsymbol{\theta})\|_{L^2(\pi_0)} \leq \psi(\ell)\,, \tag{35}$$

where the $\|\cdot\|_{L^2(\pi_0)}$ is the L^2 norm over π_0; cf. (49) in Appendix A.

Assumption 6 *There exists a constant $b_3 > 0$ independent of ℓ such that*

$$\mu^{(\ell)}_{T_\ell}(\boldsymbol{\theta}) \leq b_3\pi_0(\boldsymbol{\theta}) \tag{36}$$

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_ℓ converges in an L^2-sense to G as the level ℓ is increased. At the same time, as the level ℓ is increased and G_ℓ gets closer to G, the computational costs of evaluating G_ℓ may increase with a rate γ. 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 (Stuart, 2010, Corollary 4.9), although there a point-wise 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 5 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 (Robert and Casella, 2004).

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

$$c^*_{\mathrm{ML}}(\epsilon) \leq \frac{c_0 s^{2\gamma}}{\lambda\gamma\log(s)} \log\left(s^\alpha + (1 + s^\alpha)\left(\frac{4}{3} + \frac{b_3}{3b_1b_2}\right)\right)\left(\frac{\sqrt{3b_1b_2b_0}}{\epsilon}\right)^{2\gamma/\alpha}\,, \tag{37}$$

where the constants b_1, b_2 are independent of ϵ and given in the proof of Lemma 7 in Appendix B.

Proof By Lemma 8 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 $R_\ell \leq k_3 s^{-\alpha\ell}$ for some constant k_3 to apply

Proposition 3.

$$
\begin{aligned}
R_\ell &= \int_\Theta \left(\mu^{(\ell-1)}_{T_{\ell-1}}(\boldsymbol{\theta}) - \pi^{(\ell-1)}(\boldsymbol{\theta})\right) \log\left(\frac{\pi^{(\ell-1)}(\boldsymbol{\theta})}{\pi^{(\ell)}(\boldsymbol{\theta})}\right) \, d\boldsymbol{\theta} \\
&= \int_\Theta \left(\mu^{(\ell-1)}_{T_{\ell-1}}(\boldsymbol{\theta}) - \pi^{(\ell-1)}(\boldsymbol{\theta})\right) \log\left(\frac{Z_\ell \exp\left(-\frac{1}{2}\|\mathbf{y} - G_{\ell-1}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}{Z_{\ell-1} \exp\left(-\frac{1}{2}\|\mathbf{y} - G_{\ell}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}\right) \, \mathrm{d}\boldsymbol{\theta} \\
&= \int_\Theta \left(\mu^{(\ell-1)}_{T_{\ell-1}}(\boldsymbol{\theta}) - \pi^{(\ell-1)}(\boldsymbol{\theta})\right) \log\left(\frac{\exp\left(-\frac{1}{2}\|\mathbf{y} - G_{\ell-1}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}{\exp\left(-\frac{1}{2}\|\mathbf{y} - G_{\ell}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}\right) \, \mathrm{d}\boldsymbol{\theta},
\end{aligned} \tag{38}
$$

where the last line follows from the fact that

$$
\int_\Theta \left(\mu^{(\ell-1)}_{T_{\ell-1}}(\boldsymbol{\theta}) - \pi^{(\ell-1)}(\boldsymbol{\theta})\right) \log\left(\frac{Z_\ell}{Z_{\ell-1}}\right) \, \mathrm{d}\boldsymbol{\theta} = 0 \tag{39}
$$

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

$$
\begin{aligned}
R_\ell \leq \frac{1}{2} \int_\Theta & \left| \|\mathbf{y} - G_\ell(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G_{\ell-1}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \mu^{(\ell-1)}_{T_{\ell-1}}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\
&+ \frac{1}{2} \int_\Theta \left| \|\mathbf{y} - G_\ell(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G_{\ell-1}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \pi^{(\ell-1)}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \, .
\end{aligned} \tag{40}
$$

We have that

$$
\pi^{(\ell-1)}(\boldsymbol{\theta}) \leq \frac{1}{Z_{\ell-1}} \pi_0(\boldsymbol{\theta}) \, , \tag{41}
$$

so that when combined with Assumption 6

$$
\begin{aligned}
R_\ell &\leq \frac{1}{2} \int_\Theta \left| \|\mathbf{y} - G_\ell(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G_{\ell-1}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \mu^{(\ell-1)}_{T_{\ell-1}}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\
&\quad + \frac{1}{2} \int_\Theta \left| \|\mathbf{y} - G_\ell(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G_{\ell-1}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \pi^{(\ell-1)}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\
&\leq \frac{b_3}{2} \int_\Theta \left| \|\mathbf{y} - G_\ell(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G_{\ell-1}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \pi_0(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\
&\quad + \frac{1}{2Z_{\ell-1}} \int_\Theta \left| \|\mathbf{y} - G_\ell(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G_{\ell-1}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \pi_0(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\
&\leq \left(\frac{b_3}{2} + \frac{b_1 b_2}{2}\right) \|G_\ell - G_{\ell-1}\|_{L^2(\pi_0)},
\end{aligned} \tag{42}
$$

so that $k_3 = \left(\frac{b_3}{2} + \frac{b_1 b_2}{2}\right) b_0(1 + s^\alpha)$. Plugging in the values of k_1, k_2, and k_3 into Proposition 3 gives the result. ■

Algorithm 1 Discrete MLSVGD with adaptive stopping criterion

```
Inputs: (unnormalized) densities $\pi^{(1)}, \dots, \pi^{(L)}$, initial particles $\{\boldsymbol{\theta}_0^{[i]}\}_{i=1}^N$, step size $\delta$, tolerance $\epsilon$
Result: Particles $\{\boldsymbol{\theta}_t^{[i]}\}_{i=1}^N$
for $\ell = 1, \dots, L$ do
    repeat
        Set $s_i = \nabla \log \pi^{(\ell)}(\boldsymbol{\theta}_t^{[i]})$ for $i = 1, \dots, N$
        for $i = 1, \dots, N$ do
            $\boldsymbol{\theta}_{t+\delta}^{[i]} = \boldsymbol{\theta}_t^{[i]} + \frac{\delta}{N}\left(\sum_{j=1}^N \nabla_1 K(\boldsymbol{\theta}_t^{[j]}, \boldsymbol{\theta}_t^{[i]}) + \sum_{j=1}^N K(\boldsymbol{\theta}_t^{[j]}, \boldsymbol{\theta}_t^{[i]}) s_j\right)$
        end
        Estimate the norm of the gradient $\hat{\boldsymbol{g}}_t^{(\ell)}$ as in (43)
        Set $t \leftarrow t + \delta$
    until $\hat{\boldsymbol{g}}_t^{(\ell)} \leq \epsilon$;
end
```

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 $T_1^*, \dots, T_L^*$.

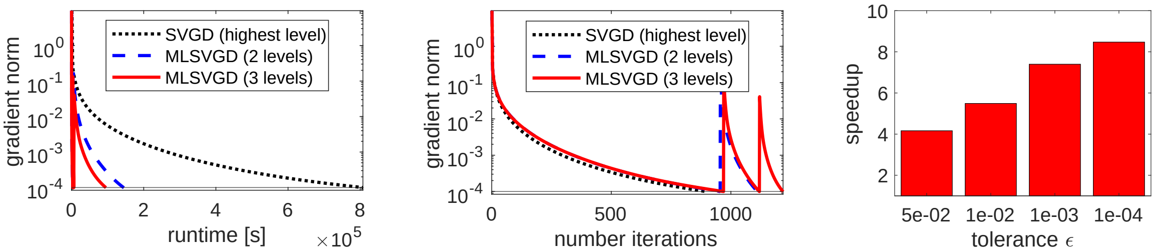
Let $\boldsymbol{g}_t^{(\ell)}$ denote the functional gradient of the KL divergence, as discussed in Section 2.1, at $\mu_t^{(\ell)}$ with target measure $\pi^{(\ell)}$. We approximate the expected norm of the gradient $\mathbb{E}_{\boldsymbol{\theta} \sim \mu_t^{(\ell)}} \left\| \boldsymbol{g}_t^{(\ell)}(\boldsymbol{\theta}) \right\|$ with the estimator

$$\hat{\boldsymbol{g}}_t^{(\ell)} = \frac{1}{N} \sum_{i=1}^N \left\| \sum_{j=1}^N \nabla_1 K(\boldsymbol{\theta}_t^{[j]}, \boldsymbol{\theta}_t^{[i]}) + \sum_{j=1}^N K(\boldsymbol{\theta}_t^{[j]}, \boldsymbol{\theta}_t^{[i]}) \nabla \log \pi^{(\ell)}(\boldsymbol{\theta}_t^{[j]}) \right\| , \tag{43}$$

where we note that each term in the sum is computed during the update (2). The adaptive stopping criterion used in Algorithm 1 is to terminate the iterations at level ℓ whenever $\hat{\boldsymbol{g}}_t^{(\ell)} \leq \epsilon$. Ideally, one would want to track the KL divergence between the SVGD approximation and the target distribution and switch to the following level once the KL divergence is below some specified threshold. However, because the normalized target density as well as the density of the SVGD approximation itself are unknown, attempting to monitor the KL divergence at each iteration is impractical. The adaptive stopping criterion based on the gradient norm, which we use, is motivated by (Duncan et al., 2019, Equation 61). It states that for small perturbations from the target density, the KL divergence between the perturbed distribution and the target distribution is asymptotically the same as the norm of the gradient squared.

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



(a) runtime, tolerance $\epsilon = 10^{-4}$ (b) iterations, tolerance $\epsilon = 10^{-4}$ (c) MLSVGD (3 levels)

Figure 2: Diffusion-reaction: MLSVGD achieves speedups because most of the iterations are on lower, cheaper levels, in contrast to SVGD that performs all iterations on the highest, most expensive level. A spike in the gradient norm indicates switching to a higher level.

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

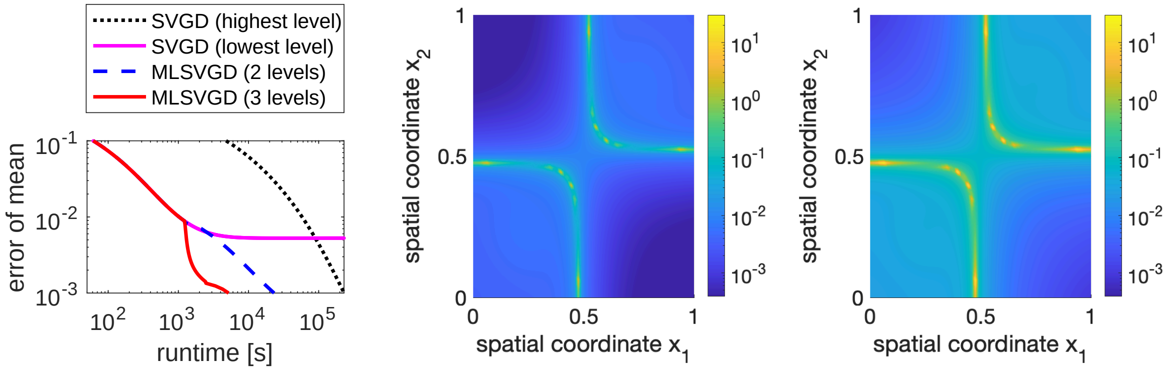
$$-\nabla^2 u(x_1, x_2; \boldsymbol{\theta}) + g(u(x_1, x_2; \boldsymbol{\theta}), \boldsymbol{\theta}) = 100 \sin(2\pi x_1) \sin(2\pi x_2)\,, \quad \mathbf{x} \in \Omega\,, \tag{44}$$

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

$$g(u(\mathbf{x}; \boldsymbol{\theta}), \boldsymbol{\theta}) = (0.1 \sin(\theta_1) + 2) \exp(-2.7\theta_1^2)(\exp(1.8\theta_2 u(\mathbf{x}; \boldsymbol{\theta})) - 1)\,.$$

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 $G_\ell(\boldsymbol{\theta}) \in \mathcal{Y}$ 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 $\boldsymbol{\theta}^* = [-\pi/4, 3]^T$ and consider the data $\mathbf{y} = G_{L+1}(\boldsymbol{\theta}^*) + \boldsymbol{e}$, where $L = 3$ (i.e., $h = 2^{-5}$) and $\boldsymbol{e}$ 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 $k(\boldsymbol{\theta}, \boldsymbol{\theta}') = \exp\big(-\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2/(2\sigma_k)\big)$ 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 ϵ. We also run MLSVGD as in Algorithm 1 with levels $\ell \in \{1,2,3\}$ and $\ell \in \{1,3\}$.



(a) error w.r.t. MCMC reference

(b) MLSVGD

(c) SVGD (same costs as (b))

Figure 3: Diffusion-reaction: MLSVGD reaches a particle mean with error 10^{-3} with respect to an MCMC reference with more than one order of magnitude speedup compared to SVGD.

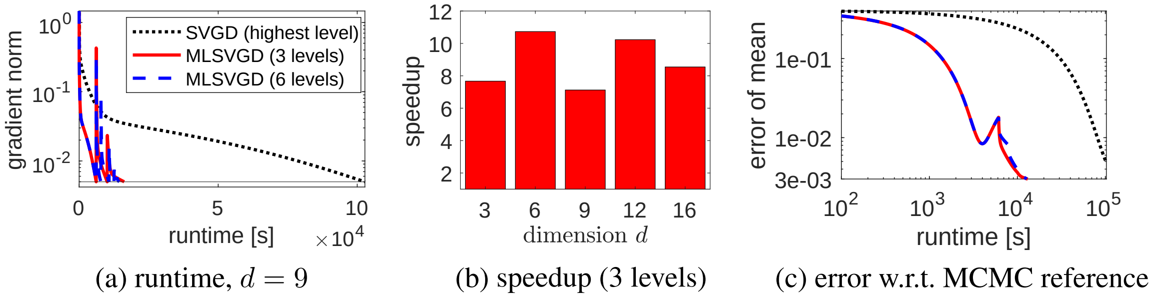
Results Figure 2 shows the decay of the estimated gradient norm (43) for SVGD and MLSVGD with two and three levels, respectively, for a tolerance $\epsilon = 10^{-4}$. While the number of total iterations over all levels in MLSVGD is higher than in SVGD, the costs per iteration are lower on lower levels 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 ($\ell \in \{1, 3\}$) achieves a slightly lower speedup than MLSVGD with 3 levels in this example. Figure 2c shows the speedup of MLSVGD with 3 levels for various tolerances. The speedup increases as the tolerance decreases. Figure 3a 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 $\ell = 1$ is fast but leads to a bias of the particle mean as indicated by the leveling off of the corresponding curve. 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

$$\partial_x^2(E(x)\partial_x^2 u(x)) = f(x) , \quad x \in \Omega , \tag{45}$$

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 PDE (45) is discretized with finite differences on a mesh of 601 equidistant grid points in Ω. The observation $\mathbf{y} \in \mathbb{R}^{41}$ is the displacement at 41 equidistant points in Ω polluted with 0.01% zero-mean Gaussian noise. We consider a smoothed piecewise constant approximation



(a) runtime, $d = 9$ (b) speedup (3 levels) (c) error w.r.t. MCMC reference

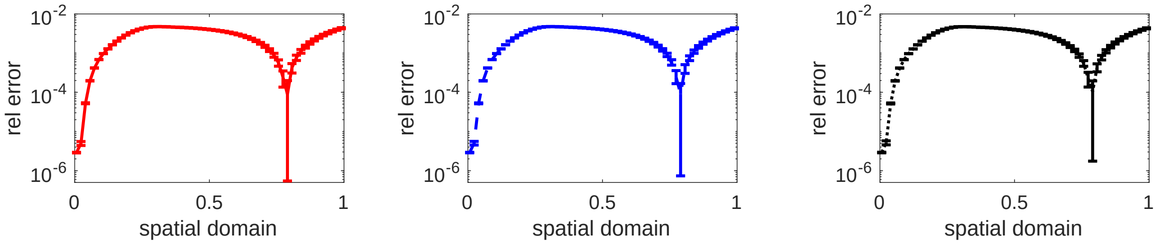
Figure 4: Euler-Bernoulli: MLSVGD achieves speedups between 6–10 in this example compared to SVGD. (Plots (c) shown for $d = 9$.)

$\hat{E}_d$ of the stiffness E that depends on $d \in \mathbb{N}$ parameters $\boldsymbol{\theta} = [\theta_1, \dots, \theta_d]^T$, cf. Appendix E. The parameter-to-observable map G_ℓ is then given by numerically solving (45) with stiffness $\hat{E}_d$ on level ℓ. The levels $\ell = 1, \dots, 6$ are corresponding to a discretization of the PDE on a mesh of $51, 101, \dots, 501$ 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 $[1, 1, \dots, 1]^T \in \mathbb{R}^d$ and diagonal covariance with 4×10^{-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 σ_k is 10^{-6} for $d = 3$ and 10^{-5} for $d \in \{6, 9\}$ and 5×10^{-5} for $d \in \{12, 16\}$. We consider MLSVGD for levels $\ell \in \{1, \dots, 6\}$ and $\ell \in \{1, 3, 6\}$. The rest of the setup is the same as in Section 6.1. Figure 4a 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 ϵ. 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 3×10^{-3} 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



(a) MLSVGD (3 levels) (b) MLSVGD (6 levels) (c) single-level SVGD

Figure 5: Euler-Bernoulli: The pointwise errors over an ensemble of inferred solutions obtained with MLSVGD (left and middle) is comparable to the errors obtained with the computationally more expensive single-level SVGD (right) in this example. The error bars show the minimum and maximum error over the ensemble. Results are shown for $d = 9$.

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.

Acknowledgments

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 under Award Number FA9550-21-1-0222 (Dr. Fariba Fahroo).

References

T. Alsup and B. Peherstorfer. Context-aware surrogate modeling for balancing approximation and sampling costs in multi-fidelity importance sampling and Bayesian inverse problems. *arXiv:2010.11708*, 2020.

D. Bakry, I. Gentil, and M. Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, 2014.

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.

W. Briggs, V. E. Henson, and S. McCormick. *A Multigrid Tutorial, Second Edition*. Society for Industrial and Applied Mathematics, second edition, 2000. doi: 10.1137/1.9780898719505.

H. Bungartz and M. Griebel. Sparse grids. *Acta Numerica*, 13:147–269, 2004. doi: 10.1017/S0962492904000182.

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.

S. Chewi, T. Le 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.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020.

J. Andrés Christen and C. Fox. Markov chain Monte Carlo using an approximation. *Journal of Computational and Graphical Statistics*, 14(4):795–810, 2005. ISSN 1061-8600.

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.

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.

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. doi: 10.1137/130915005.

A. Duncan, N. Nuesken, and L. Szpruch. On the geometry of Stein variational gradient descent. *arXiv:1912.00894*, 2019.

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.

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

A. Gregory, C. J. Cotter, and S. Reich. Multilevel ensemble transform particle filtering. *SIAM Journal on Scientific Computing*, 38(3):A1317–A1338, 2016. doi: 10.1137/15M1038232.

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

H. Haario, M. Laine, A. Mira, and E. Saksman. DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354, December 2006. ISSN 0960-3174. doi: 10.1007/s11222-006-9438-0.

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

J. Han and Q. Liu. Stein variational adaptive importance sampling. In A. Ihler, editor, *Conference on Uncertainty in Artificial Intelligence*. Curran Associates, Inc., 2017.

S. Heinrich. Multilevel Monte Carlo methods. In Svetozar Margenov, Jerzy Waśniewski, and Plamen Yalamov, editors, *Large-Scale Scientific Computing*, pages 58–67, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

V. Ha Hoang, C. Schwab, and A. M. Stuart. Complexity analysis of accelerated MCMC methods for Bayesian inversion. *Inverse Problems*, 29(8):085010, jul 2013. doi: 10.1088/0266-5611/29/8/085010.

H. Hoel, K. Law, and R. Tempone. Multilevel ensemble Kalman filtering. *SIAM Journal on Numerical Analysis*, 54(3):1813–1839, 2016.

A. Jasra, K. Kamatani, K. Law, and Y. Zhou. Multilevel particle filters. *SIAM Journal on Numerical Analysis*, 55(6):3068–3096, 2017. doi: 10.1137/17M1111553.

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.

A. Korba, A. Salim, M. Arbel, G. Luise, and A. Gretton. A non-asymptotic analysis for Stein variational gradient descent. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4672–4682. Curran Associates, Inc., 2020.

J. Latz, I. Papaioannou, and E. Ullmann. Multilevel sequential^2 Monte Carlo for Bayesian inverse problems. *Journal of Computational Physics*, 368:154 – 178, 2018.

C. Liu, J. Zhuo, P. Cheng, R. Zhang, and J. Zhu. Understanding and accelerating particle-based variational inference. In Kamalika Chaudhuri and Ruslan 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.

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.

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.

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.

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

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

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

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

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, 2018a.

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

R. Ranganath, S. Gerrish, and D. Blei. Black Box Variational Inference. In Samuel Kaski and Jukka 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.

D. Rezende and S. Mohamed. Variational inference with normalizing flows. In Francis Bach and David 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.

C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

A.M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

F. Wagner, J. Latz, I. Papaioannou, and E. Ullmann. Multilevel sequential importance sampling for rare event estimation. *SIAM Journal on Scientific Computing*, 42(4):A2062–A2087, 2020.

C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026, 2019. doi: 10.1109/TPAMI.2018.2889774.

Appendix A. Metrics and divergences and other definitions

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

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

The Kullback-Leibler (KL) divergence from μ to η is defined as

$$\mathrm{KL}(\mu||\eta) = \int_{\mathbb{R}^d} \mu(\boldsymbol{\theta}) \log\left(\frac{\mu(\boldsymbol{\theta})}{\eta(\boldsymbol{\theta})}\right) \,\mathrm{d}\boldsymbol{\theta}\,. \tag{47}$$

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

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

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

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

Appendix B. Lemma 7 and proof

Lemma 7 *If Assumption 5 holds, there exists a constant $C > 0$ such that for all $1 \leq \ell_1, \ell_2 \leq \infty$ sufficiently large*

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

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

We note that this proof closely mirrors the proofs of Lemma 4.2 and 4.3 in (Marzouk and Xiu, 2009), 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 $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^d$ and symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ we have

$$\begin{aligned} \|\boldsymbol{u} - \boldsymbol{w}\|^2_{\mathbf{A}} - \|\boldsymbol{v} - \boldsymbol{w}\|^2_{\mathbf{A}} &= \|(\boldsymbol{u} - \boldsymbol{v}) + (\boldsymbol{v} - \boldsymbol{w})\|^2_{\mathbf{A}} - \|\boldsymbol{v} - \boldsymbol{w}\|^2_{\mathbf{A}} \\ &= \langle (\boldsymbol{u} - \boldsymbol{v}),\ \mathbf{A}(\boldsymbol{u} - \boldsymbol{v})\rangle + 2\langle (\boldsymbol{u} - \boldsymbol{v}),\ \mathbf{A}(\boldsymbol{v} - \boldsymbol{w})\rangle \\ &= \langle (\boldsymbol{u} - \boldsymbol{v}),\ \mathbf{A}(\boldsymbol{u} + \boldsymbol{v} - 2\boldsymbol{w})\rangle \\ &\leq \|\boldsymbol{u} - \boldsymbol{v}\| \cdot \|\mathbf{A}(\boldsymbol{u} + \boldsymbol{v} - 2\boldsymbol{w})\|\,, \end{aligned} \tag{51}$$

with the last line following from the Cauchy-Schwarz inequality. Applying this bound with $\boldsymbol{u} = G_1(\boldsymbol{\theta})$, $\boldsymbol{v} = G_2(\boldsymbol{\theta})$, $\boldsymbol{w} = \mathbf{y}$, and $\mathbf{A} = \boldsymbol{\Gamma}^{-1}$ gives

$$\begin{aligned} &\int_\Theta \left| \|\mathbf{y} - G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G_2(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \pi_0(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta} \\ &\leq \int_\Theta \|G_1(\boldsymbol{\theta}) - G_2(\boldsymbol{\theta})\| \cdot \|\boldsymbol{\Gamma}^{-1}(2\mathbf{y} - G_1(\boldsymbol{\theta}) - G_2(\boldsymbol{\theta}))\| \pi_0(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta} \\ &\leq \|G_1 - G_2\|_{L^2(\pi_0)} \cdot \|\boldsymbol{\Gamma}^{-1}(2\mathbf{y} - G_1 - G_2)\|_{L^2(\pi_0)}\,, \end{aligned} \tag{52}$$

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)

$$\begin{aligned}
\mathrm{KL}(\pi_1 \,||\, \pi_2) &= \int_\Theta \pi_1(\boldsymbol{\theta}) \log\left(\frac{\pi_1(\boldsymbol{\theta})}{\pi_2(\boldsymbol{\theta})}\right)\, \mathrm{d}\boldsymbol{\theta} \\
&= \int_\Theta \pi_1(\boldsymbol{\theta}) \log\left(\frac{Z_2 \exp\left(-\frac{1}{2}\|\mathbf{y}-G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}{Z_1 \exp\left(-\frac{1}{2}\|\mathbf{y}-G_2(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}\right)\, \mathrm{d}\boldsymbol{\theta} \\
&= \log\left(\frac{Z_2}{Z_1}\right) + \int_\Theta \pi_1(\boldsymbol{\theta}) \log\left(\frac{\exp\left(-\frac{1}{2}\|\mathbf{y}-G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}{\exp\left(-\frac{1}{2}\|\mathbf{y}-G_2(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}\right)\, \mathrm{d}\boldsymbol{\theta} \\
&\leq \log\left(\frac{Z_2}{Z_1}\right) + \frac{1}{2Z_1}\int_\Theta \left|\|\mathbf{y}-G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y}-G_2(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right| \pi_0(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\
&\leq \left|\log\left(\frac{Z_2}{Z_1}\right)\right| + \frac{1}{2Z_1}\|G_1 - G_2\|_{L^2(\pi_0)} \cdot \|\boldsymbol{\Gamma}^{-1}(2\mathbf{y} - G_1 - G_2)\|_{L^2(\pi_0)}\,,
\end{aligned} \tag{53}$$

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

$$\exp\left(-\frac{1}{2}\|\mathbf{y}-G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right) \leq 1\,. \tag{54}$$

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)

$$\begin{aligned}
|Z_1 - Z_2| &= \left|\int_\Theta \left\{\exp\left(-\frac{1}{2}\|\mathbf{y}-G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right) - \exp\left(-\frac{1}{2}\|\mathbf{y}-G_2(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)\right\} \pi_0(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}\right| \\
&\leq \int_\Theta \left|\exp\left(-\frac{1}{2}\|\mathbf{y}-G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right) - \exp\left(-\frac{1}{2}\|\mathbf{y}-G_2(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)\right| \pi_0(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\
&\leq \frac{1}{2}\int_\Theta \left|\|\mathbf{y}-G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y}-G_2(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right| \pi_0(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\
&\leq \frac{1}{2}\|G_1 - G_2\|_{L^2(\pi_0)} \cdot \|\boldsymbol{\Gamma}^{-1}(2\mathbf{y} - G_1 - G_2)\|_{L^2(\pi_0)}\,.
\end{aligned} \tag{55}$$

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

$$\begin{aligned}
\|\boldsymbol{\Gamma}^{-1}(2\mathbf{y} - G_1 - G_2)\|_{L^2(\pi_0)} &\leq 2\|\boldsymbol{\Gamma}^{-1}\mathbf{y}\|_{L^2(\pi_0)} + \|\boldsymbol{\Gamma}^{-1}(G_1 + G_2)\|_{L^2(\pi_0)} \\
&\leq 2\|\boldsymbol{\Gamma}^{-1}\mathbf{y}\|_{L^2(\pi_0)} + 2\|\boldsymbol{\Gamma}^{-1}G\|_{L^2(\pi_0)} + \|\boldsymbol{\Gamma}^{-1}(G_1 + G_2 - 2G)\|_{L^2(\pi_0)} \\
&\leq 2\|\boldsymbol{\Gamma}^{-1}\mathbf{y}\|_{L^2(\pi_0)} + 2\|\boldsymbol{\Gamma}^{-1}G\|_{L^2(\pi_0)} \\
&\quad + \frac{1}{\gamma_{\min}}\|G_1 - G\|_{L^2(\pi_0)} + \frac{1}{\gamma_{\min}}\|G_2 - G\|_{L^2(\pi_0)}\,.
\end{aligned} \tag{56}$$

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 ℓ_1 and ℓ_2. Therefore, there exists a constant $b_1 > 0$ independent of ℓ such that

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

Combining Equations (55) and (57) yields

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

The ratio of the normalizing constants can be written

$$\left|\frac{Z_2}{Z_1} - 1\right| = \frac{1}{Z_1}\left|Z_1 - Z_2\right|\,, \tag{59}$$

so the logarithm can be bounded as

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

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

$$\left|\log\left(\frac{Z_2}{Z_1}\right)\right| \leq \max\left\{\frac{\frac{|Z_2-Z_1|}{Z_1}}{1 - \frac{|Z_2-Z_1|}{Z_1}},\ \frac{|Z_2 - Z_1|}{Z_1}\right\} \leq \frac{|Z_1 - Z_2|}{Z_1 - |Z_1 - Z_2|}\,. \tag{61}$$

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

$$Z_1^{-1} \leq \sup_{\ell\geq 1} Z_\ell^{-1} \leq b_2\,. \tag{62}$$

Moreover, for all ℓ_1, ℓ_2 sufficiently large $|Z_1 - Z_2| \leq b_2^{-1}/2$. Using the bound gives

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

Combining Equations (53), (57), (58), (62), and (63) gives

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

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

Appendix C. Lemma 8 and proof

Lemma 8 *If Assumption 5 holds with $\psi(\ell) = b_0 s^{-\alpha\ell}$, then Assumptions 2, 4 also hold with the same rate α.*

Proof Let $\ell_1 = \ell$ and $\ell_2 = \infty$, so that by Lemma 7 in Appendix B we immediately have

$$\mathrm{KL}(\pi^{(\ell)} \mid\mid \pi) \leq C\|G_\ell - G\|_{L^2(\pi_0)} \leq C\psi(\ell) = Cb_0 s^{-\alpha\ell}, \tag{65}$$

so that $k_1 = Cb_0$. Moreover, setting $\ell_1 = \ell - 1$ and $\ell_2 = \ell$ and using the triangle inequality gives

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

Thus,

$$\mathrm{KL}(\pi^{(\ell-1)} \mid\mid \pi^{(\ell)}) \leq C\left(1 + \frac{\psi(\ell - 1)}{\psi(\ell)}\right)\psi(\ell) \leq Cb_0\left(1 + s^\alpha\right)s^{-\alpha\ell}\,, \tag{67}$$

so that $k_2 = Cb_0\left(1 + s^\alpha\right)$. ■

Appendix D. A triangle-like inequality for the KL divergence

Let ρ_0, ρ_1, ρ_2 be three probability distributions on Θ. We have that

$$\begin{aligned}
\mathrm{KL}(\rho_0 \,||\, \rho_2) &= \int_\Theta \rho_0(\boldsymbol{\theta}) \log\left(\frac{\rho_0(\boldsymbol{\theta})}{\rho_2(\boldsymbol{\theta})}\right) \,\mathrm{d}\boldsymbol{\theta} \\
&= \int_\Theta \rho_0(\boldsymbol{\theta}) \log\left(\frac{\rho_0(\boldsymbol{\theta})\rho_1(\boldsymbol{\theta})}{\rho_1(\boldsymbol{\theta})\rho_2(\boldsymbol{\theta})}\right) \,\mathrm{d}\boldsymbol{\theta} \\
&= \int_\Theta \rho_0(\boldsymbol{\theta}) \log\left(\frac{\rho_0(\boldsymbol{\theta})}{\rho_1(\boldsymbol{\theta})}\right) \,\mathrm{d}\boldsymbol{\theta} + \int_\Theta \rho_0(\boldsymbol{\theta}) \log\left(\frac{\rho_1(\boldsymbol{\theta})}{\rho_2(\boldsymbol{\theta})}\right) \,\mathrm{d}\boldsymbol{\theta} \\
&= \mathrm{KL}(\rho_0 \,||\, \rho_1) + \mathrm{KL}(\rho_1 \,||\, \rho_2) + \int_\Theta (\rho_0(\boldsymbol{\theta}) - \rho_1(\boldsymbol{\theta})) \log\left(\frac{\rho_1(\boldsymbol{\theta})}{\rho_2(\boldsymbol{\theta})}\right) \,\mathrm{d}\boldsymbol{\theta}\,,
\end{aligned} \tag{68}$$

cf. the inequality given in (Marzouk and Xiu, 2009). We refer to this third term in the last line as the remainder term.

Appendix E. Details about numerical experiments

E.1. General

The step size δ and kernel bandwidth h was chosen via a manual process so that SVGD on the highest level numerically converged. The same δ 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 (Haario et al., 2001, 2006) 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 $\bar{\boldsymbol{\theta}}$ of the parameter. The error reported in Figure 3a and Figure 4c is $\frac{1}{10}\sum_{i=1}^{10} \|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(i)}\|_2$, where $\boldsymbol{\theta}^{(i)}$ is the mean of (ML)SVGD particles of the i-th replicate.

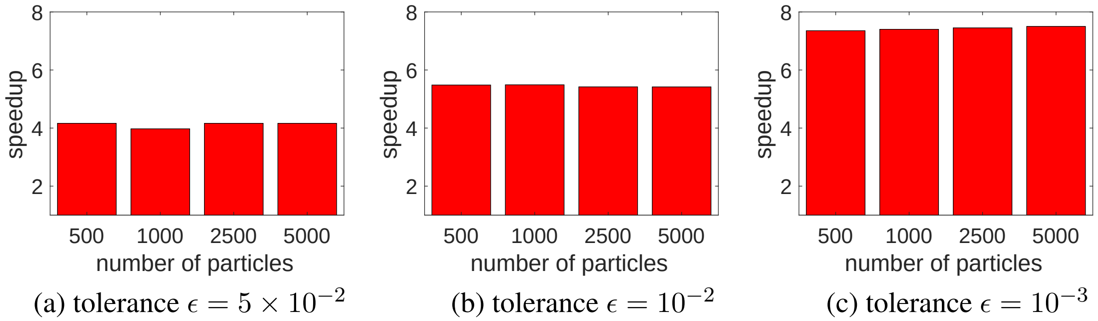
E.2. Diffusion equation with nonlinear reaction term

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

$$u(0) = 0\,, \quad \frac{\partial}{\partial x} u\bigg|_{x=0} = 0\,, \quad \frac{\partial^3}{\partial x^3} u\bigg|_{x=1} = 0\,, \quad \frac{\partial^3}{\partial x^3} u\bigg|_{x=1} = 0\,.$$



(a) tolerance $\epsilon = 5 \times 10^{-2}$ (b) tolerance $\epsilon = 10^{-2}$ (c) tolerance $\epsilon = 10^{-3}$

Figure 6: Diffusion-reaction: The cost of MLSVGD scales with the number of particles as the cost of SVGD, which means that the speedups that MLSVGD obtains compared to SVGD in this example remain roughly the same for different number of particles.

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

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

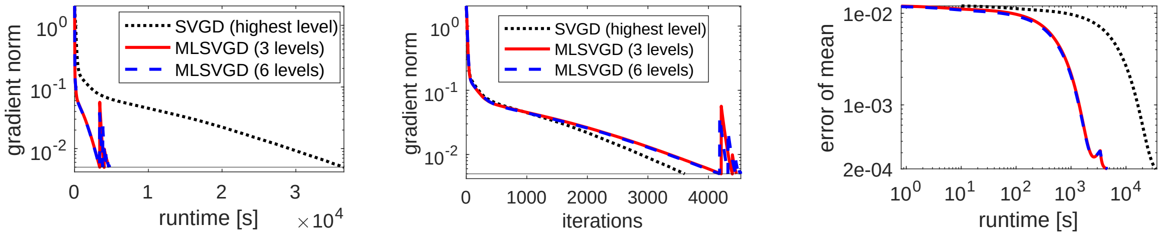
such that there is a smooth transition from 0 to 1 at α. For $k > 1$, let $\alpha_1, \dots, \alpha_{k+1}$ be $k + 1$ equidistant points in Ω. Let $\mathbb{R}_+ = \{z \in \mathbb{R} : z > 0\}$ and consider the parameter $\boldsymbol{\theta} = [\theta_1, \dots, \theta_k]^T \in \mathbb{R}_+^k$. Define the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

$$\hat{E}_i(x, \theta_i) = (1 - I(x, \alpha_i))\hat{E}_i(x, \theta_{i-1})$$

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

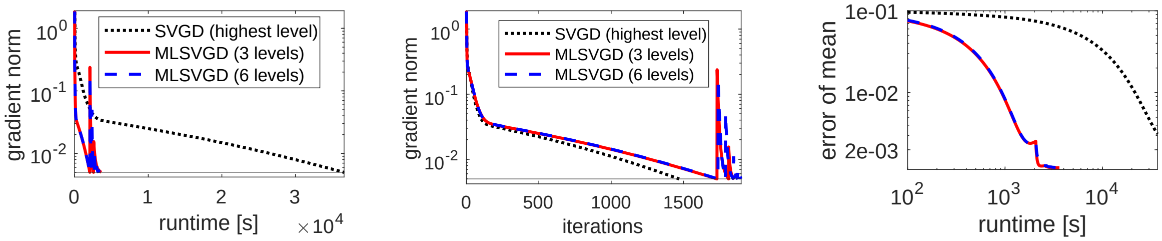
Additional plots for $d \in \{3, 6, 12, 16\}$. Figure 7–10 show the analogous results to Figure 4 for dimension $d \in \{3, 6, 12, 16\}$, respectively. Figure 11 shows the analogous results to Figure 5. The behavior of MLSVGD compared to SVGD is qualitatively the same as for dimension $d = 9$.

[1] https://github.com/g2s3-2018/labs



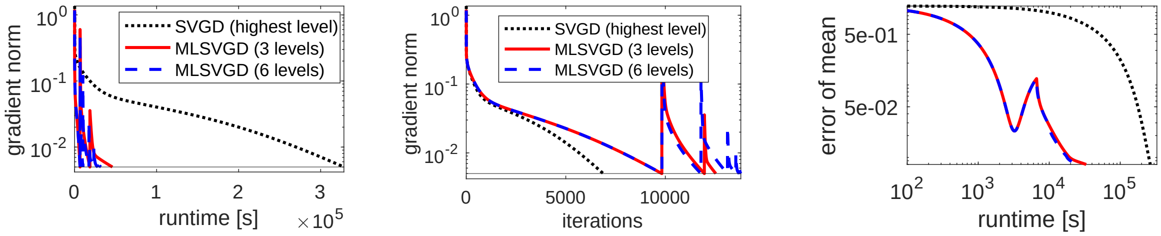
(a) runtime (b) iterations (c) comparison to MCMC reference

Figure 7: Euler-Bernoulli beam: Results of MLSVGD for dimension $d = 3$.



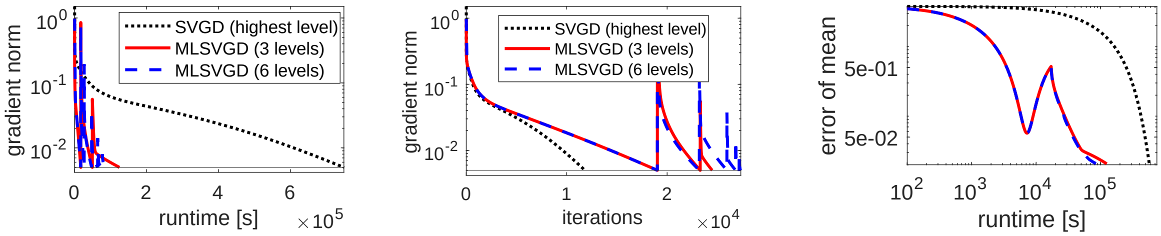
(a) runtime (b) iterations (c) comparison to MCMC reference

Figure 8: Euler-Bernoulli beam: Results of MLSVGD for dimension $d = 6$.



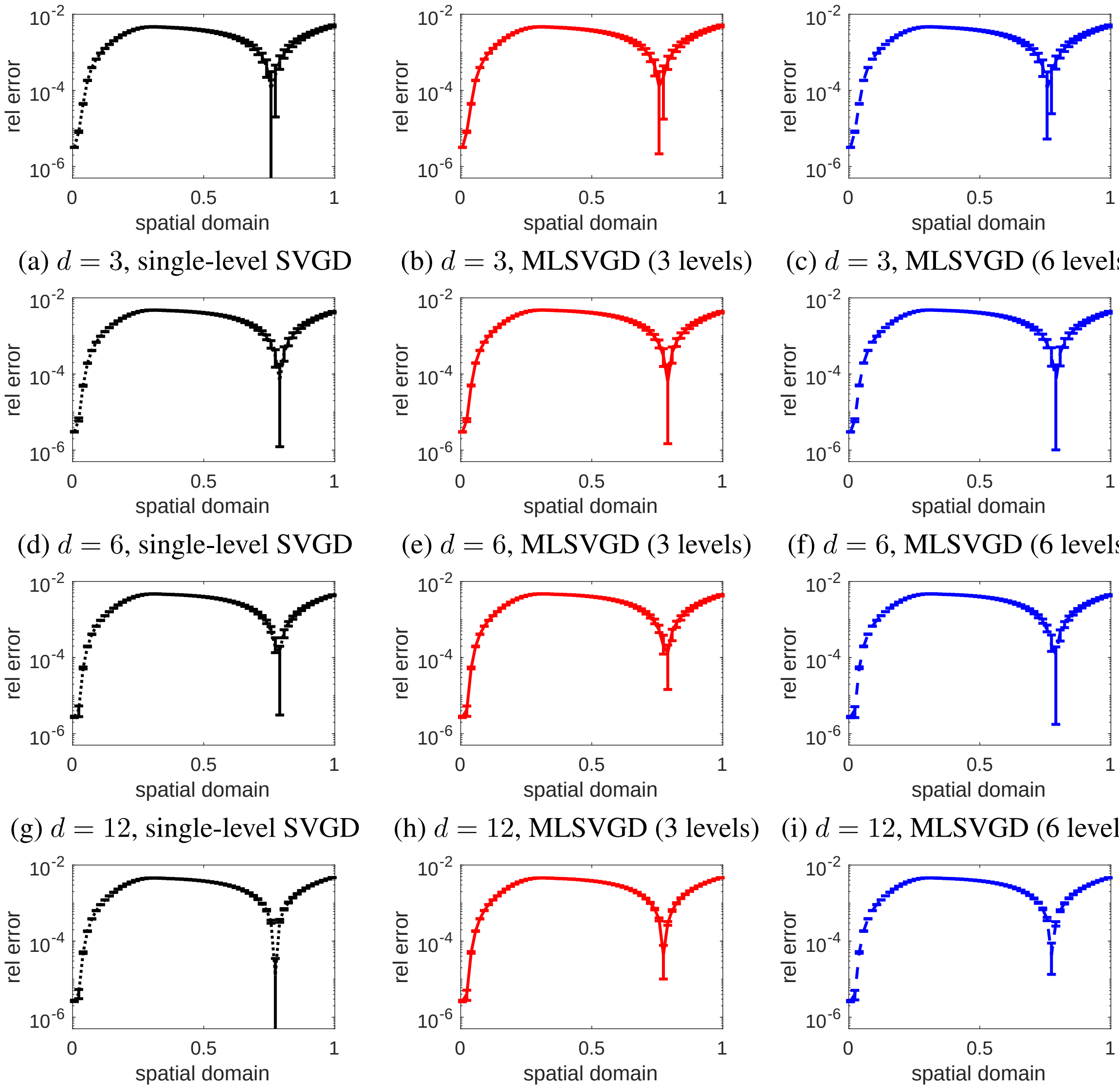
(a) runtime (b) iterations (c) comparison to MCMC reference

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



(a) runtime (b) iterations (c) comparison to MCMC reference

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



(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)

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