Trading off deterministic approximations and sampling in multifidelity Bayesian inference

by

Terrence Alsup

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Mathematics

New York University

January, 2023

Professor Benjamin Peherstorfer

Dedication

To my friends and family.

Acknowledgements

Before anyone else I would like to acknowledge my incredible advisor Ben Peherstorfer. Over the time that I've known Ben, his guidance and support have helped me grow tremendously as a researcher. His constructive feedback and suggestions have taught me valuable skills and practices that will stay with me for a long time. I also want to thank my wonderful committee members, not only for taking the time to carefully review my work, but each one has helped shape my research and interests in unique ways. Georg's class on Inverse Problems, Esteban's classes on Optimal Transport, and Jonathan's class on Monte Carlo methods all contributed to my current research interests. Similarly, Youssef's survey paper on measure transport was fundamental to helping my understanding of the topic.

There are many others who I would like to extend my gratitude to. Tommie Catanach at Sandia National Laboratories for her mentorship during Summer 2021 and beyond with insightful career advice. Everyone in the SciML group, who gave informative presentations on exciting new research fronts as well as the RTG in Mathematical Modeling and Simulation. My collaborators: Luca Venturi, Tucker Hartland, Noemi Petra, and Aimee Maurais for the unique perspectives each one brought to our projects. Finally, I want to emphasize my appreciation for all of my friends and family that have helped and supported me along the way. I could not have done this without them.

Abstract

Bayesian inference is a ubiquitous and flexible tool for updating a belief (i.e., learning) about a quantity of interest given observed data, which ultimately can be used to inform upstream decision-making. In particular, Bayesian inverse problems allow one to learn from data through the lens of physics-based models, typically given in the form of a parameter-to-observable map based on a system of partial differential equations, by prescribing a posterior probability distribution that reflects prior information about the parameters as well as the observed data. The computational task underlying Bayesian inference is to approximate the posterior distribution through sampling in order to compute expectations and to quantify uncertainties of the unknown parameters, requiring many evaluations of an expensive high-fidelity physics-based model. For this purpose, multifidelity methods present an attractive avenue for reducing the computational burden of Bayesian inference by leveraging low-cost surrogate models to speedup computations while making limited recourse to expensive high-fidelity models to establish accuracy guarantees of the final inferred quantities. Because the surrogate and high-fidelity models are used together, poor approximations by the surrogate models can be compensated with more frequent recourse to the high-fidelity model during sampling. Thus, multifidelity methods give rise to a trade-off between investing computational resources needed to learn a good deterministic approximation and the computational resources needed for sampling with respect to the high-fidelity posterior distribution. We introduce two methods: context-aware importance sampling and multilevel Stein variational

gradient descent. The first method selects a single optimal surrogate model to derive the approximation, while the second uses a hierarchy of surrogate models in a sequential fashion to derive increasingly accurate approximations. For both methods, we show through both theoretical cost complexity bounds and numerical examples that these approaches achieve up to multiple orders of magnitude in speedup when compared to their traditional counterparts.

Contents

List of Figures

List of Appendices

Chapter 1

Introduction

1.1 Multifidelity methods for outer loop applications

1.1.1 Outer-loop applications

Computational models form the bedrock of many scientific and engineering applications. For a wide range of problems, high-fidelity models are needed to sufficiently resolve the underlying physical system but require significant computational costs to evaluate. Moreover, many applications have an outer-loop structure that involves repeatedly evaluating the high-fidelity model at different inputs, potentially being intractable when the number of high-fidelity model evaluations needed is high. Several prominent examples of outer-loop applications include the following:

- **Inference**: Inferring a quantity of interest, e.g. the mean or variance, from a probability distribution via Monte Carlo methods typically requires repeated evaluations of the log density or its gradient to draw samples.

- **Optimization**: Many optimization methods iteratively update a set of design variables to minimize an objective function that depends on the high-fidelity model by evaluating its derivatives.

- **Data assimilation**: Filtering-based methods that integrate observational data into existing models typically alternate between a prediction step where the model is evaluated and an update step where the model parameters are adjusted to reflect the new data.

- **Control**: Controlling a system to a desired state with feedback requires monitoring the system by evaluating the model for the system and adjusting the control variables accordingly.

In each of these cases, when the number of model evaluations is prohibitive one may resort to replacing the high-fidelity model with a low-fidelity or surrogate model that is cheaper to evaluate. While this may make the application in question more tractable, the surrogate model may fail to capture important or fine behavior of the high-fidelity model resulting in a biased outer-loop result depending on accuracy of the surrogate model. In contrast, multifidelity methods [Peherstorfer et al., 2018d] leverage both the high-fidelity model and the available surrogate models to reduce the computational costs of the outer-loop application while maintaining accuracy of the final outer-loop result. Effective multifidelity methods delegate the bulk of the computation to the low-fidelity models which may be orders of magnitude faster to evaluate than the high-fidelity model while making limited recourse to the high-fidelity model to ensure accuracy. In this thesis, we narrow our focus specifically to the task of inference.

1.1.2 Types of Surrogate Models

There are a variety of surrogate models each with different accuracy guarantees and costs. When the underlying physical system is modeled by a partial differential equation (PDE) or ordinary differential equation (ODE), coarse grid surrogate models can be obtained by solving the PDE on a coarser mesh. Coarse grid surrogate models typically enjoy approximation guarantees from numerical analysis and are straightforward to implement by changing the number of grid points or using built-in mesh refinement functions in the same code used to compute the high-fidelity model [Trottenberg et al., 2001, Giles, 2008]. Data-fit models [Forrester and Keane, 2009, Rasmussen and Williams, 2016, Hastie et al., 2001, Swischuk et al., 2019, Qian et al., 2020] are another class of surrogate models where machine learning techniques leverage large amounts of data to learn a good approximation to the high-fidelity model. These models are often black-box and non-intrusive making them suitable to almost any high-fidelity model, but often require lots of data with limited accuracy guarantees. Recently, scientific machine learning has focused on using expressive deep neural networks as surrogate models to approximate the solution to high-dimensional PDEs [Raissi et al., 2019, Bruna et al., 2022, Dissanayake and Phan-Thien, 1994, Sirignano and Spiliopoulos, 2018] that are out of reach for traditional grid-based methods. Simplified-physics models [Ng and Willcox, 2016, Majda and Gershgorin, 2010, Cao et al., 2011, Konrad et al., 2021] are another class of surrogate models that aim to capture the important physical behavior, often by considering linearized models, but require an understanding of the underlying process. These surrogate models range in difficulty of their implementation. In some cases, closed-form solutions may be available while in others one may need to write completely new code for to solve the simplified model. Yet another class of surrogate models are projection-based models such as system-theoretic model reduction [Antoulas, 2005, Antoulas et al., 2020] and reduced basis and POD methods [Quarteroni et al., 2011, Benner et al., 2015, Hesthaven

et al., 2016, Chen et al., 2017], which capture the dominant singular values of the solution. Often these models have approximation guarantees but are typically intrusive. An exception are non-intrusive and data-driven model reduction methods [Peherstorfer and Willcox, 2016, Benner et al., 2020, Ionita and Antoulas, 2014, Qian et al., 2020] which do not require access to the high-fidelity model, making them amenable to legacy code bases, for example.

1.1.3 Multilevel methods

Multifidelity methods can leverage heterogeneous surrogate models to speed up computation of the outer-loop result. In contrast, multilevel methods use a hierarchy of low-fidelity models with increasing accuracy and costs. Typically, the multilevel hierarchy is obtained through coarse grid discretizations where the level corresponds to the number of grid points or degrees of freedom in the discretized system. In these cases, one can obtain rates on the error of the surrogate models as well as rates on the cost of the surrogate models in terms of the level. These rates are often essential for determining how much computation should be invested at each level. For example, in multilevel Monte Carlo (MLMC) [Cliffe et al., 2011, Giles, 2008, Teckentrup et al., 2013, Haji-Ali et al., 2016] the rates are used to determine the number of samples to take at each level. Because of their increasing accuracy, multilevel methods have also been used successfully for hierarchical preconditioning in optimization [Weissmann et al., 2022, Li et al., 2021] where lower levels are used to find good initializations for the more expensive higher levels, see also the works [Gorodetsky et al., 2020b, Robinson et al., 2006, Lam et al., 2015] for multifidelity optimization when there is no clear hierarchy. This is akin to classical multigrid methods [Briggs et al., 2000, Hackbush, 1985, Trottenberg et al., 2001] which utilize coarse discretizations to speed up solving a system on finer discretizations.

1.2 Statistical inference

In the remainder of this thesis we will consider the task of inferring a quantity of interest of the form

$$\mathbb{E}_\pi[f] = \int_\Theta f(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{1.1}$$

from a target distribution π over $\Theta \subset \mathbb{R}^d$ with f being an integrable test function. Note that in the following we write π to denote both the target distribution and its density function. In low dimensions, typically less than three, and when f is smooth, one may use a quadrature rule to efficiently compute the integral (1.1). For moderate dimensions quasi-Monte Carlo [Morokoff and Caflisch, 1995] and sparse grid techniques [Bungartz and Griebel, 2004, Nobile et al., 2008] are available but quickly become limited as the dimension increases. In higher dimensions, however, one must resort to Monte Carlo methods. The standard Monte Carlo estimator draws independent and identically distributed (i.i.d.) samples $\{\boldsymbol{\theta}^{[i]}\}_{i=1}^N$ from the target distribution π and computes

$$\hat{f}_N^{\mathrm{MC}} = \frac{1}{N}\sum_{i=1}^N f(\boldsymbol{\theta}^{[i]}) \tag{1.2}$$

to estimate the quantity of interest (1.1).

Two scenarios that give rise to the inference problem of computing (1.1) are forward uncertainty propagation and inverse uncertainty quantification. In forward uncertainty propagation we can sample directly from the target distribution π and the test function f depends on a high-fidelity model so that computing the estimator (1.2) gives rise to an outer-loop. In this scenario we are equivalently estimating the mean of the pushforward distribution $f_\#\pi$ (hence "propagation") such that $f_\#\pi(A) = \pi(f^{-1}(A))$ for any Borel set $A \subset \Theta$. A notable example is rare event estimation where $f(\boldsymbol{\theta}) = \mathbf{1}\{G(\boldsymbol{\theta}) \geq t\}$, G is the high-fidelity model, and $t \in \mathbb{R}$ is a threshold. For inverse uncertainty quantification, or Bayesian inverse

problems, π is a posterior distribution where the likelihood depends on given observed data and a high-fidelity model. In this setting π may only be known up to a normalizing constant and cannot be sampled from directly, hence the outer loop application may be to use Markov chain Monte Carlo (MCMC) or possibly variational inference to draw approximate samples from π. We review Bayesian inverse problems in more detail in Section 1.3.

1.2.1 Trading off deterministic approximations and sampling for rejection sampling

When π cannot be sampled from directly, we may instead draw samples $\{\boldsymbol{\theta}^{[i]}\}_{i=1}^{N} \sim \mu$ from a proposal distribution μ that approximates π and then apply an accept/reject step [Robert and Casella, 2004] to obtain independent samples from π. Set

$$D_{\pi,\mu} = \max_{\boldsymbol{\theta}\in\Theta} \quad \frac{\pi(\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\,, \tag{1.3}$$

which necessarily satisfies $D_{\pi,\mu} \geq 1$, by the fact that μ and π must both integrate to one, with $D_{\pi,\mu} = 1$ if and only if $\pi = \mu$. Rejection sampling, summarized in Algorithm 1, proceeds by first sampling a proposal $\boldsymbol{\theta}'$ from the proposal distribution μ as well as an independent uniform random variable $U \sim \text{Uniform}[0,1]$. The proposed sample is accepted if $U \leq \pi(\boldsymbol{\theta}')/(D_{\pi,\mu}\mu(\boldsymbol{\theta}'))$ and rejected otherwise. After N samples have been accepted, the rejection sampling estimator becomes

$$\hat{f}_N^{\text{RS}} = \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}^{[i]})\,. \tag{1.4}$$

The constant $D_{\pi,\mu}$ quantifies how close the approximating density μ is to the target density π with larger values indicating lower acceptance rates and poorer approximations. The expected number of trials until a proposal is accepted is $D_{\pi,\mu}$ and therefore we expect

Algorithm 1: Rejection sampling with an auxiliary distribution

Inputs: Target density π, auxiliary density μ and constant $D_{\pi,\mu}$;
Result: Samples $\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[N]}$
Initialize the set of target samples $S = \{\}$ and index $i = 1$;
while $|S| < N$ **do**
  Sample $\boldsymbol{\theta}' \sim \mu$ and $U \sim \text{Uniform}[0, 1]$;
  Compute acceptance probability $\alpha = \frac{\pi(\boldsymbol{\theta}')}{D_{\pi,\mu}\mu(\boldsymbol{\theta}')}$;
  if $U \leq \alpha$ **then**
    Set $\boldsymbol{\theta}^{[i]} = \boldsymbol{\theta}'$ and $S = S \cup \boldsymbol{\theta}^{[i]}$;
    $i + 1 \leftarrow i$;
  end
end

$ND_{\pi,\mu}$ total evaluations of the target density to compute the estimator (1.4). Learning an accurate approximation μ will often require information from the target density π and incur computational costs, for example by evaluating the high-fidelity model that π depends on. Thus, there is a trade-off of computational resources between learning an accurate *deterministic approximation* and the *sampling effort* required.

In a multifidelity setting where a surrogate density $\pi^{(\ell)}$ is available, one may instead learn the approximation μ using information about $\pi^{(\ell)}$ to reduce the cost of constructing the deterministic approximation at the expense of requiring potentially more sampling effort through a larger constant $D_{\pi,\mu}$. In general we have

$$D_{\pi,\mu} \leq \left(\max_{\boldsymbol{\theta}\in\Theta} \quad \frac{\pi(\boldsymbol{\theta})}{\pi^{(\ell)}(\boldsymbol{\theta})}\right)\left(\max_{\boldsymbol{\theta}\in\Theta} \quad \frac{\pi^{(\ell)}(\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\right) = D_{\pi,\pi^{(\ell)}} D_{\pi^{(\ell)},\mu}\,, \tag{1.5}$$

which shows that the sampling effort required to compute (1.4) depends on both the fidelity ℓ of the surrogate density as well as how well μ approximates the surrogate density $\pi^{(\ell)}$. Moreover, if there are many surrogate models with a hierarchical structure, one can use them sequentially to cheaply learn an approximation μ, which we consider in more detail in Chapter 3. Decompositions such as (1.5) are commonplace in the analysis of multifidelity

methods for inference and typically are written in terms of a metric or divergence on the space of probability measures. For example, the work [Marzouk and Xiu, 2009] considers the KL divergence in the context of stochastic collocation. In Chapter 2 we consider importance sampling where the sampling effort is controlled by the χ^2 divergence and in Chapter 3 we consider Stein variational gradient descent [Liu and Wang, 2016] where the sampling effort is determined by the KL divergence.

1.2.2 Multifidelity methods for sampling

We give a brief overview of related literature on multifidelity and multilevel methods for sampling.

Forward UQ

First there are methods geared towards forward uncertainty propagation where one is able to directly sample from the target distribution but must evaluate the high-fidelity model at the samples. These methods are primarily variance reduction methods which attempt to limit the number of high-fidelity samples needed as much as possible. The two common approaches are control variates and importance sampling. The classical approach is multilevel Monte Carlo [Giles, 2008, Cliffe et al., 2011] which relies on correlations between surrogate models at consecutive levels to obtain a control variates estimator with reduced variance. Other multifidelity control variates estimators can leverage surrogate models when there is no clear hierarchy [Peherstorfer, 2019, Maurais, 2022, Peherstorfer et al., 2016b, Peherstorfer et al., 2018a, Peherstorfer et al., 2018b, Kramer et al., 2019] and can be thought of as regressions on the low-fidelity models [Schaden and Ullmann, 2020, Gorodetsky et al., 2020a]. On the other hand, multifidelity importance sampling [Peherstorfer et al., 2016a, Peherstorfer et al., 2017] methods use surrogate models to approximate an optimal biasing density that minimizes the variance. Importance sampling methods are particularly successful for rare event estimation [Li and Xiu, 2010, Li et al., 2011, Chen and Quarteroni, 2013] and multiple

surrogate models may be used in a sequential fashion as in the multifidelity cross-entropy method [Peherstorfer et al., 2018c] or multilevel sequential Monte Carlo [Wagner et al., 2020].

Inverse UQ

When the target distribution cannot be directly sampled one must instead draw approximate samples, for example from an auxiliary or proposal distribution as discussed in Section 1.2.1. A general approach is transport-map based [Moselhy and Marzouk, 2012, Marzouk et al., 2016] and variational approximations [Ranganath et al., 2014, Zhang et al., 2019] where one fits a tractable distribution to the target distribution that can then be sampled directly. Normalizing flows [Rezende and Mohamed, 2015, Tabak and Turner, 2012, Tabak and Vanden-Eijnden, 2010] are a particular example where an efficient neural network architecture is used to parameterize the density. Because variational approximations are typically fit by minimizing the Kullback-Leibler (KL) divergence from the target, multilevel optimization techniques such as [Li et al., 2021, Weissmann et al., 2022] may be used to reduce computational costs. Stein variational gradient descent (SVGD) [Liu and Wang, 2016] and its multilevel extension [Alsup et al., 2021] is a nonparametric form of variational inference that relies on an ensemble of particles instead. We present these methods in detail in Chapter 3. While variational approximations have the advantage of being able to draw independent samples, they are asymptotically biased.

Monte Carlo methods, on the other hand, particularly Markov chain Monte Carlo (MCMC), produce consistent estimates at the cost of the samples being correlated with potentially long autocorrelation times [Liu, 2004]. In general, successful MCMC methods with faster decaying autocorrelation learn a good proposal to generate better samples such as affine invariant ensemble samplers [Goodman and Weare, 2010, Leimkuhler et al., 2018] and parametric-based approaches that are trained adaptively [Gabrié et al., 2022]. Moreover, there are 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]. Variational approximations may also be used in tandem with MCMC to serve as better proposal densities [Parno and Marzouk, 2018, Gabrié et al., 2022] with a multifidelity extension in [Peherstorfer and Marzouk, 2019]. In addition to better proposal densities for MCMC, variational approximations may serve as better biasing densities for importance sampling [Alsup and Peherstorfer, 2022], which we discuss in detail in Chapter 2.

Finally, there are particle-based methods that iteratively evolve an ensemble of particles to approximate the target distribution such as ensemble Kalman filters [Iglesias et al., 2013, Schillings and Stuart, 2017] and sequential Monte Carlo [Liu, 2004]. Multilevel particle filters [Jasra et al., 2017], multilevel sequential Monte Carlo [Beskos et al., 2017] methods, and multilevel ensemble Kalman filtering [Hoel et al., 2016] update high and low-fidelity particles concurrently and rely on telescoping sums of correlated differences between successive levels similar to multilevel Monte Carlo [Cliffe et al., 2011]. Additionally one can use transport maps to extend to nonlinear filtering [Gregory et al., 2016]. In contrast, the multilevel sequential Monte Carlo methods presented in [Latz et al., 2018, Wagner et al., 2020] use a sequence of distributions to approximate the target distribution similar to multilevel SVGD discussed in Chapter 3.

1.3 Bayesian inverse problems

We provide a brief overview of Bayesian inverse problems [Stuart, 2010, Kaipio and Somersalo, 2007, Sullivan, 2015] which serve as a prototypical example where the target distribution depends on a high-fidelity model and cannot be sampled but has a hierarchy of surrogate models available which can be leveraged in a multifidelity or multilevel method. Classical inverse problems seek to recover a set of model parameters $\boldsymbol{\theta}^\star \in \Theta \subset \mathbb{R}^d$ given measured data $\mathbf{y} \in \mathbb{R}^q$. The model parameters and the measured data are related through a parameter-to-

observable map $G : \Theta \to \mathbb{R}^q$ that describes the underlying physical system by

$$\mathbf{y} = G(\boldsymbol{\theta}^\star) + \boldsymbol{\eta}\,, \tag{1.6}$$

where $\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{\Gamma})$ is noise that corrupts the observations. The noise model (1.6) defines a likelihood function $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$ for the parameter given the data

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{y}) = \frac{1}{(2\pi)^{q/2}|\boldsymbol{\Gamma}|^{1/2}} \exp\left(-\frac{1}{2}\left\|\mathbf{y} - G(\boldsymbol{\theta})\right\|^2_{\boldsymbol{\Gamma}^{-1}}\right)\,, \tag{1.7}$$

where $\|\mathbf{v}\|^2_{\boldsymbol{\Gamma}^{-1}} = \left\langle \mathbf{v},\ \boldsymbol{\Gamma}^{-1}\mathbf{v}\right\rangle$. In general the likelihood function does not define a probability distribution over the parameter as it may not integrate to 1. However, if additional information encoded in a prior distribution π_0 is available, then Bayes' rule gives a posterior distribution over the parameters

$$\pi(\boldsymbol{\theta}) = \frac{1}{Z}\exp\left(-\frac{1}{2}\left\|\mathbf{y} - G(\boldsymbol{\theta})\right\|^2_{\boldsymbol{\Gamma}^{-1}}\right)\pi_0(\boldsymbol{\theta})\,, \tag{1.8}$$

where Z is a normalization constant

$$Z = \int_\Theta \exp\left(-\frac{1}{2}\left\|\mathbf{y} - G(\boldsymbol{\theta})\right\|^2_{\boldsymbol{\Gamma}^{-1}}\right)\pi_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{1.9}$$

to ensure that the posterior is a probability distribution. When the parameter-to-observable map G is nonlinear, the posterior π may be intractable to sample from its normalizing constant Z unknown.

1.3.1 Multifidelity methods for Bayesian inverse problems

For many scientific and engineering applications, the parameter-to-observable map G depends on the solution of an underlying PDE or system of PDEs that describes a physical

system. In these cases, the parameter-to-observable map can be written as the composition $G = \mathcal{B}^{\text{obs}} \circ F$ of a solution operator $F : \Theta \to \mathcal{U}$ that maps the parameters to the solution of the PDE in a function space $\mathcal{U}$ and an observation operator $\mathcal{B}^{\text{obs}} : \mathcal{U} \to \mathbb{R}^q$, which is typically a linear functional of the solution e.g. pointwise observations at specified points. Moreover, when the system of PDEs cannot be solved exactly, and thus we cannot directly evaluate G, we must resort to a numerical method that discretizes the underlying PDE problem to approximately evaluate the solution operator F. Let $F^{(\ell)}$ denote such an approximation with $G^{(\ell)} = \mathcal{B}^{\text{obs}} \circ F^{(\ell)}$ as the corresponding surrogate parameter-to-observable map. The index ℓ denotes the fidelity of the surrogate model and for example may correspond to the number of elements in a finite element [Brenner and Scott, 2008] approximation, the number of grid points in finite differences [LeVeque, 2007], the number of times steps for an ordinary differential equation [LeVeque, 2007], the number of terms in a Karhunen-Loève expansion [Sullivan, 2015], and others. Larger fidelities ℓ correspond to more accurate the approximation $G^{(\ell)}$ of G with surrogate posterior densities defined as

$$\pi^{(\ell)}(\boldsymbol{\theta}) = \frac{1}{Z_\ell} \exp\left(-\frac{1}{2}\left\|\mathbf{y} - G^{(\ell)}(\boldsymbol{\theta})\right\|^2_{\boldsymbol{\Gamma}^{-1}}\right) \pi_0(\boldsymbol{\theta})\,, \tag{1.10}$$

and Z_ℓ defined analogously as Z (1.9). Although in this thesis we only consider finite dimensional parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ one may also consider the case where the parameter corresponds to a general Banach space, for example when the parameter is a field or function. For discussion of infinite dimensional Bayesian inverse problems we refer to the works [Stuart, 2010, Sullivan, 2015]. In certain cases we may approximate the solution of an infinite dimensional inverse problem in a low-dimensional function space by defining an interpolation operator $\mathcal{I}^{\text{int}} : \Theta \to \mathcal{U}_0$ where $\mathcal{U}_0$ is a function space containing the infinite dimensional parameter. We present several concrete examples of this in both Chapters 2 and 3.

Remark 1. In Chapter 2 we deviate slightly from this notation and instead let $G^{(h)}$ denote

the surrogate parameter-to-observable map (model) at fidelity h. In this setting, small $h > 0$ corresponds to more accurate higher-fidelity models with larger h corresponding to lower-fidelity models. Although the difference is primarily notational, in this setting specifically we consider a continuum of fidelities h that may denote, for example, the mesh width or time step size in the solution of an ODE. The rest of the set up is the same, e.g. $\pi^{(h)}$ corresponds to the surrogate densities.

1.3.2 Linear Bayesian inverse problems

In general, the posterior distribution cannot be sampled directly and we must use one of the methods discussed in Section 1.2.2 to perform inference. An exception to this is when the parameter-to-observable model G is linear with a Gaussian noise model (1.6) and prior $\pi_0 = N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ so that the posterior π becomes a Gaussian as well. For linear inverse problems we may express the parameter-to-observable map G as a matrix $\mathbf{G}$ so that $G(\boldsymbol{\theta}) = \mathbf{G}\boldsymbol{\theta}$. The posterior mean is given as a weighted average of the prior mean $\boldsymbol{\mu}_0$ and the data $\mathbf{y}$

$$\boldsymbol{\mu} = \left(\boldsymbol{\Sigma}_0^{-1} + \mathbf{G}^\top \boldsymbol{\Gamma}^{-1}\mathbf{G}\right)^{-1} \left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \mathbf{G}^\top \boldsymbol{\Gamma}^{-1}\mathbf{y}\right) , \tag{1.11}$$

and the posterior covariance is a weighted harmonic average of the prior covariance $\boldsymbol{\Sigma}_0$ and the covariance $\mathbf{G}^\top \boldsymbol{\Gamma}^{-1}\mathbf{G}$

$$\boldsymbol{\Sigma} = \left(\boldsymbol{\Sigma}_0^{-1} + \mathbf{G}^\top \boldsymbol{\Gamma}^{-1}\mathbf{G}\right)^{-1} . \tag{1.12}$$

Linear Bayesian inverse problems can serve as a helpful tool for understanding the behavior of posterior distributions even in the nonlinear setting. For example, from (1.12) we see that if $\boldsymbol{\Gamma} = \gamma\mathbf{I}$ for some constant $\gamma > 0$, then

$$\boldsymbol{\Sigma} = \gamma \left(\gamma\boldsymbol{\Sigma}_0^{-1} + \mathbf{G}^\top \mathbf{G}\right)^{-1} ,$$

asymptotically approaches $\gamma \mathbf{G}^\top \mathbf{G}$ as the noise level $\gamma \to 0$. The Bernstein-von Mises theorem [van der Vaart, 1998] implies that the posterior behaves asymptotically as Gaussian as the noise level approaches zero. Small noise levels γ may correspond to either informative or large quantities of data. If the matrix $\mathbf{G}$ is full rank, then the posterior will concentrate around the data $\mathbf{y}$ as $\gamma \to 0$. However if $\mathbf{G}$ is not full rank, the inverse problem is ill-conditioned and the posterior will not concentrate around the data even as the noise level shrinks. Finally, because the posteriors in linear Bayesian inverse problems are Gaussian we can often compute exactly the probability distances and divergences between the surrogate densities $\pi^{(\ell)}$ and the target density π which are needed for determining the sampling effort as in Section 1.2.1.

1.4 Context-aware learning: Trade-offs in multifidelity inference

1.4.1 Challenges of multifidelity inference

Multifidelity methods provide a principled way for combining both high-fidelity models and available low-fidelity models to speed up outer-loop applications while maintaining accuracy of the outer-loop result. However, the manner in which low-fidelity models are constructed, see Section 1.1.2, is often divorced from the outer-loop application itself. This leads to an inefficiency where sub-optimal surrogate models may be used to assist in the outer-loop task. On one extreme, the outer-loop application may be insensitive to the accuracy of the low-fidelity resulting in wasted computational resources additionally evaluating the low-fidelity model. On the other extreme, the outer-loop application may require a very accurate surrogate model to be effective and offers no improvement over using the high-fidelity model alone. With regards to inference, poor approximating distributions derived from sub-optimal

surrogate densities can result in considerable sampling effort. Thus, there arises a need to understand the trade-off between the accuracy of the deterministic approximation and the effort required by the sampling procedure in order to derive more efficient approximations. Furthermore, for multilevel methods where a hierarchy of increasingly accurate surrogate models are available, one must determine how much sampling effort to expend on each level.

1.4.2 Contributions

While traditional surrogate models are often constructed in ignorance of the particular outer-loop application, e.g. inference, context-aware surrogate models [Peherstorfer, 2019, Farcas, 2020, Alsup and Peherstorfer, 2022, Werner and Peherstorfer, 2022, Shyamkumar et al., 2022, Farcas et al., 2022] specifically take the outer-loop application (i.e., the *context*) into consideration to further reduce the computational cost.

- We introduce context-aware importance sampling, published in [Alsup and Peherstorfer, 2022], that adaptively selects a surrogate model based on the outer-loop application as opposed to traditional multifidelity importance sampling where a static surrogate model is chosen.

- We present multilevel Stein variational gradient descent, which we introduced and developed in our works [Alsup et al., 2021, Alsup et al., 2022] respectively, that utilizes a hierarchy of surrogate models to sequentially derive good initializations for SVGD thereby reducing the total computational cost.

- For both context-aware importance sampling and MLSVGD we derive cost-complexity bounds as well as demonstrate speedups numerically on a suite a different problems. Both results indicate speedups of the proposed methods over their traditional counterparts.

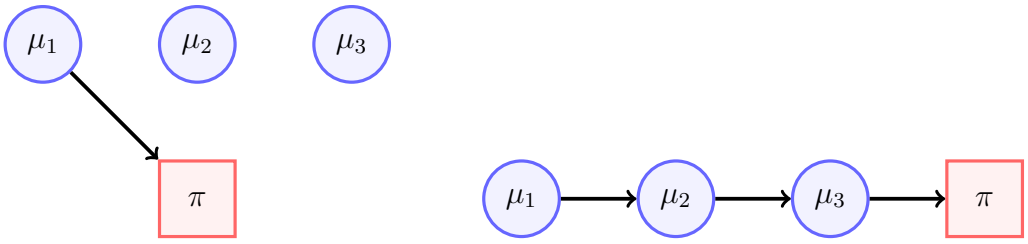


Figure 1.1: (**Left**) Single-level multifidelity methods, such as context-aware importance sampling, select a single optimal surrogate model. (**Right**) Hierarchical multifidelity methods, such as MLSVGD, chain together increasingly accurate approximations.

Previously, context-aware methods for inference relied on control variates [Peherstorfer, 2019]. In Chapter 2, we present context-aware importance sampling (CAIS), which we published in [Alsup and Peherstorfer, 2022], where the selected surrogate model is used to derive a biasing density for importance sampling. Here we derive an optimization problem that describes the trade-off between the fidelity of the surrogate model used to learn the biasing density and the sampling effort required to re-weight the samples with respect to the high-fidelity model. We obtain theoretical cost complexity bounds for the CAIS estimator as well as the traditional multifidelity importance sampling (MFIS) estimator [Peherstorfer et al., 2016a], where the fidelity is fixed, to show that the CAIS estimator achieves the same error tolerance at a reduced cost. Further, both estimators are tested on three different Bayesian inverse problems where we observe an order of magnitude speedup for the CAIS estimator coinciding with our theoretical results.

While context-aware importance sampling selects a single optimal approximation for importance sampling, in Chapter 3 we present a multilevel extension to SVGD, which we published in [Alsup et al., 2021] and provided further analysis for in [Alsup et al., 2022], that chains together a hierarchy of approximations to reduce the computational cost, c.f. Figure 1.1. To do so we use information on the rates at which the surrogate models converge as well as the rate at which SVGD converges in order to prescribe the sampling effort needed at each level. We again derive cost complexity bounds for both SVGD and MLSVGD to

show that MLSVGD enjoys a theoretical speedup and demonstrate speedups up to one order of magnitude in three separate numerical examples. In particular, we demonstrate that MLSVGD can efficiently infer glacier ice flow for the Haut Glacier d'Arolla [Alsup et al., 2022].

Chapter 2

Context-aware importance sampling

In this chapter we present our published work [Alsup and Peherstorfer, 2022] which derives an optimal trade-off between surrogate-model fidelity and computational costs for multifidelity importance sampling (MFIS) [Peherstorfer et al., 2016a]. To derive the trade-off we develop bounds on the error of the MFIS estimator that depend on the surrogate-model fidelity that we then combine with existing bounds on the error of general importance sampling estimators [Chatterjee and Diaconis, 2018, Agapiou et al., 2017, Sanz-Alonso, 2018]. As in Section 1.2.1, these error bounds take the form of a probability divergence between the target distribution and the biasing distribution, which we use to separate the statistical error due to sampling and the deterministic approximation error due to the quality of the biasing density. Here the biasing density is taken to be the Laplace approximation of the surrogate density due to its favorable approximation guarantees and tractability to both compute and sample from. Note that there is a large body of work on adaptive importance sampling that studies minimizing the χ^2 divergence to derive an optimal biasing density [Al-Qaq et al., 1995, Ryu and Boyd, 2014, Akyildiz and Míguez, 2021], but these works do not consider the cost of surrogate models during training as we do here.

2.1 Preliminaries

2.1.1 Notation and problem setting

Let $(\Theta, \mathcal{B}(\Theta), \pi)$ denote a probability space where $\Theta = \mathbb{R}^d$ is the domain for parameters $\boldsymbol{\theta}$, $\mathcal{B}(\Theta)$ is the Borel σ-algebra of Θ, and π is a probability distribution on Θ. Let π admit a density function with respect to the Lebesgue measure on $\mathbb{R}^d$ and refer to both the distribution and the density function as $\pi : \Theta \to \mathbb{R}$. In many applications, particularly Bayesian inference, c.f. Section 1.3, the density π may only be evaluated up to a normalizing factor

$$\pi = \frac{1}{Z}\tilde{\pi}, \quad Z = \int_\Theta \tilde{\pi}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \,,$$

where $\tilde{\pi} \geq 0$ is the unnormalized density and $Z \in (0, \infty)$ is the normalizing constant. In the following, we consider situations where the density π and the unnormalized density $\tilde{\pi}$ are expensive to evaluate. The task at hand is to compute quantities of interest with respect to the target distribution π which take the form of expectations

$$\mathbb{E}_\pi[f] = \int_\Theta f(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}, \tag{2.1}$$

where f is a bounded measurable test function, i.e., $\|f\|_{L^\infty} < \infty$ where $\|f\|_{L^\infty} = \text{ess sup}_{\boldsymbol{\theta}\in\Theta}|f(\boldsymbol{\theta})|$ under the measure π.

2.1.2 Importance sampling

Let μ be another probability distribution on the Borel space $(\Theta, \mathcal{B}(\Theta))$ that also admits a density function with respect to the Lebesgue measure on $\mathbb{R}^d$. Again let μ refer to both the probability distribution and the density function with respect to the Lebesgue measure. Moreover, assume that π is absolutely continuous with respect to μ, so that for any Borel set

$B \in \mathcal{B}(\Theta)$ such that $\mu(B) = 0$, $\pi(B) = 0$ as well. If sampling directly from π is impossible then one may instead estimate the quantity of interest (2.1) through importance sampling with μ as the biasing density. Draw N independent and identically distributed samples $\{\boldsymbol{\theta}^{[i]}\}_{i=1}^{N} \overset{\text{i.i.d.}}{\sim} \mu$ and compute the weights

$$w(\boldsymbol{\theta}^{[i]}) = \frac{\pi(\boldsymbol{\theta}^{[i]})}{\mu(\boldsymbol{\theta}^{[i]})}, \quad i = 1, \ldots, N\,. \tag{2.2}$$

The importance sampling estimator is given by

$$\hat{f}_N^{\text{IS}} = \frac{1}{N} \sum_{i=1}^{N} w(\boldsymbol{\theta}^{[i]}) f(\boldsymbol{\theta}^{[i]})\,. \tag{2.3}$$

In the case where the normalizing constant Z, and hence the exact density π, is unknown, then self-normalized importance sampling can be used to estimate the expectation (2.1). Draw N independent and identically distributed samples $\{\boldsymbol{\theta}^{[i]}\}_{i=1}^{N} \overset{\text{i.i.d.}}{\sim} \mu$ from the biasing distribution μ and re-weight them with the target distribution π to obtain the self-normalized importance sampling estimator

$$\hat{f}_N^{\text{SNIS}} = \frac{\sum_{i=1}^{N} f(\boldsymbol{\theta}^{[i]}) \tilde{w}(\boldsymbol{\theta}^{[i]})}{\sum_{i=1}^{N} \tilde{w}(\boldsymbol{\theta}^{[i]})} \tag{2.4}$$

of $\mathbb{E}_\pi[f]$, where the importance weights $\tilde{w}(\boldsymbol{\theta}^{[i]})$ are now given by evaluating the unnormalized density ratio

$$\tilde{w}(\boldsymbol{\theta}^{[i]}) = \frac{\tilde{\pi}(\boldsymbol{\theta}^{[i]})}{\mu(\boldsymbol{\theta}^{[i]})} \tag{2.5}$$

at the samples $\boldsymbol{\theta}^{[i]}$. If all $\tilde{w}(\boldsymbol{\theta}^{[i]}) = 0$, then we define $\hat{f}_N^{\text{SNIS}} = 0$. The estimator (2.4) is a consistent estimator of $\mathbb{E}_\pi[f]$ as the sample size $N \to \infty$.

2.2 Error of the importance sampling estimator

The standard importance sampling estimator (2.3) is unbiased and therefore the mean-squared error (MSE) of the estimator is equal to its variance

$$\begin{aligned}\mathrm{MSE}[\hat{f}_N^{\mathrm{IS}}] &= \mathbb{E}\left[\left(\hat{f}_N^{\mathrm{IS}} - \mathbb{E}_\pi[f]\right)^2\right] \\ &= \mathrm{Var}\left[\hat{f}_N^{\mathrm{IS}}\right] \\ &= \frac{1}{N}\mathrm{Var}_\mu\left[\frac{\pi(\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} f(\boldsymbol{\theta})\right]\end{aligned}$$

Since the test function f is bounded we can bound

$$\frac{1}{N}\mathrm{Var}_\mu\left[\frac{\pi(\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} f(\boldsymbol{\theta})\right] \leq \frac{\|f\|_{L^\infty}^2}{N}\mathbb{E}_\mu\left[\left(\frac{\pi(\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\right)^2\right].$$

Note that because $\mathbb{E}_\mu[\pi(\boldsymbol{\theta})/\mu(\boldsymbol{\theta})] = 1$, the variance is exactly the χ^2 divergence of the target density π to the biasing density μ defined as

$$\chi^2(\pi \,||\, \mu) = \mathrm{Var}_\mu\left[\frac{\pi(\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\right] = \mathbb{E}_\mu\left[\left(\frac{\pi(\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\right)^2\right] - 1 = \int_\Theta \frac{\pi(\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} - 1\,. \tag{2.6}$$

Therefore, the mean-squared error of the importance sampling estimator (2.3) is bounded by

$$\mathrm{MSE}[\hat{f}_N^{\mathrm{IS}}] \leq \frac{\|f\|_{L^\infty}^2}{N}\left(\chi^2(\pi \,||\, \mu) + 1\right). \tag{2.7}$$

Similarly, if only the unnormalized density $\tilde{\pi}$ is available, then [Agapiou et al., 2017, Theorem 2.1] gives the following bound on the MSE of the self-normalized importance sampling estimator (2.4)

$$\mathrm{MSE}[\hat{f}_N^{\mathrm{SNIS}}] \leq \frac{4\,\|f\|_{L^\infty}^2}{N}\left(\chi^2(\pi \,||\, \mu) + 1\right). \tag{2.8}$$

Note that because the upper bounds on the right-hand-sides of (2.7) and (2.8) only depend on the test function through its norm $\|f\|_{L^\infty}$, we immediately obtain a uniform bound on the MSE over any test functions f with $\|f\|_{L^\infty} \leq M$ for some $M < \infty$. Because f is bounded, we trivially have that

$$(f - \mathbb{E}_\pi[f])^2 \leq 4 \|f\|_{L^\infty}^2 \,,$$

so that the bounds (2.7) and (2.8) are only useful if the sample size $N \geq \chi^2(\pi \,||\, \mu) + 1$ is sufficiently large and motivates the definition of the effective sample size

$$N_{\text{eff}} = \frac{N}{\chi^2(\pi \,||\, \mu) + 1} \,. \tag{2.9}$$

The effective sample size (2.9) corresponds to the number of i.i.d. samples from the target distribution π that are needed to achieve an equivalent mean-squared error as the importance sampling estimator (2.3). A large χ^2 divergence of the target π from the biasing density μ results in a lower effective sample size which motivates learning a biasing density μ with a small χ^2 divergence.

Note that our formulation of importance sampling is slightly different from the typical setting where importance sampling is used as variance reduction to estimate the quantity of interest (2.1) of a specific fixed test function f, as in the case of rare event estimation for example. In the case where f is fixed, the optimal biasing density that minimizes the MSE takes the form

$$\mu^\star(\boldsymbol{\theta}) = \frac{|f(\boldsymbol{\theta})| \, \pi(\boldsymbol{\theta})}{\int_\Theta |f(\boldsymbol{\theta}')| \, \pi(\boldsymbol{\theta}') \mathrm{d}\boldsymbol{\theta}'} \,, \tag{2.10}$$

and for $f > 0$ gives zero MSE. Learning the optimal biasing density (2.10) for a fixed test function f is a separate objective from what we consider in this work since we seek a biasing density to approximate π directly, as opposed to $|f| \, \pi$. This is again reflected in the effective sample size $N_{\text{eff}} \leq N$ which is never greater than the actual sample size N if we had drawn

i.i.d. samples from π directly.

2.2.1 Learning a biasing density

To determine an appropriate biasing density μ that results in a large effective sample size (2.9) one must use information about the unnormalized target density $\tilde{\pi}$. For example, in adaptive importance sampling one may consider a parametric family of biasing densities $\{\mu_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathcal{A}\}$ such as a Gaussian or mixture of Gaussians, transport maps, or normalizing flows [Tabak and Turner, 2012, Rezende and Mohamed, 2015, Moselhy and Marzouk, 2012] and then optimize over the parameters $\boldsymbol{\alpha}$ to obtain an optimal biasing density

$$\boldsymbol{\alpha}^* = \arg\min_{\alpha \in \mathcal{A}} \chi^2 \left(\pi \,||\, \mu_{\boldsymbol{\alpha}}\right) . \tag{2.11}$$

The practicality of the adaptive importance sampling approach depends on how tractable the optimization problem (2.11) is. For example [Ryu and Boyd, 2014] consider only exponential families for the biasing density, guaranteeing that the problem (2.11) is convex, while others may consider instead minimizing the KL divergence to fit transport maps [Moselhy and Marzouk, 2012] or normalizing flows [Tabak and Turner, 2012, Rezende and Mohamed, 2015], or as in the cross-entropy method [Peherstorfer et al., 2018c]. Moreover, while Gaussian mixture models, transport maps, and normalizing flows offer greater flexibility to more closely match the target density, they may be more computationally challenging to fit and lack approximation guarantees. Another approach, which we explore in Section 2.3.3 is when the biasing density μ is the Laplace approximation to π, which requires optimizing the log density $\log \pi$ and computing the Hessian at the optimal point.

2.2.2 Multifidelity importance sampling

When the high-fidelity target density π is computationally expensive to evaluate, the costs of learning an appropriate biasing density can be prohibitive. Instead for many applications it is beneficial to replace the high-fidelity target π with a low-fidelity surrogate density that is computationally cheaper. Let $(\pi^{(h)})_{h>0}$ denote a sequence of low-fidelity probability distributions that approximate π, where the index h denotes the fidelity of the approximation. Again we assume each $\pi^{(h)}$ admits a density with respect to the Lebesgue measure and refer to $\pi^{(h)}$ as both the density function and the distribution and that the unnormalized low-fidelity densities may be written

$$\pi^{(h)} = \frac{1}{Z_h}\tilde{\pi}^{(h)}, \quad Z_h = \int_\Theta \tilde{\pi}^{(h)}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\,.$$

Let the low-fidelity densities converge pointwise to the high-fidelity density so that for each $\boldsymbol{\theta} \in \Theta$, $\pi^{(h)}(\boldsymbol{\theta}) \to \pi(\boldsymbol{\theta})$ as $h \to 0$. Define the cost of evaluating the unnormalized high-fidelity density $\tilde{\pi}(\boldsymbol{\theta})$ at any point $\boldsymbol{\theta} \in \Theta$ to be $C^{\text{high}} > 0$ and the cost of evaluating the unnormalized low-fidelity density $\pi^{(h)}$ to be given by $c(h)$, where $c : (0,\infty) \to [0,\infty)$. Multifidelity importance sampling (MFIS) [Peherstorfer et al., 2016a] replaces the unnormalized high-fidelity density $\tilde{\pi}$ with an unnormalized low-fidelity density $\tilde{\pi}^{(h)}$ to learn a biasing density μ_h for importance sampling. The MFIS estimator for (2.1) is

$$\hat{f}^{\text{MFIS}}_{h,N} = \frac{\sum_{i=1}^N \tilde{w}_h(\boldsymbol{\theta}^{[i]})f(\boldsymbol{\theta}^{[i]})}{\sum_{i=1}^N \tilde{w}_h(\boldsymbol{\theta}^{[i]})}, \quad \{\boldsymbol{\theta}^{[i]}\} \overset{\text{i.i.d.}}{\sim} \mu_h\,, \tag{2.12}$$

with the importance weights given by

$$\tilde{w}_h(\boldsymbol{\theta}^{[i]}) = \frac{\tilde{\pi}(\boldsymbol{\theta}^{[i]})}{\mu_h(\boldsymbol{\theta}^{[i]})}\,. \tag{2.13}$$

Because the unnormalized low-fidelity density $\tilde{\pi}^{(h)}$ is only evaluated when deriving the biasing density μ_h and not when computing estimator (2.12), the MFIS estimator (2.12) is consistent with respect to the quantity of interest $\mathbb{E}_\pi[f]$ as $N \to \infty$. Moreover, because the biasing density μ_h is learned from only using the unnormalized low-fidelity density $\tilde{\pi}^{(h)}$, and not the particular test function f, the biasing density may be recycled for many different test functions. The bound (2.8) shows that the mean-squared error of the MFIS estimator (2.12) depends on the effective sample size (2.9) resulting from using the biasing density μ_h, and therefore the number of samples required to achieve a predetermined error tolerance depends directly on the fidelity h of the low-fidelity density.

2.2.3 Problem formulation

Given a low-fidelity surrogate density $\pi^{(h)}$, estimating a quantity of interest $\mathbb{E}_\pi[f]$ using the MFIS estimator (2.12) is a two-step procedure. First, one must learn a suitable biasing density μ_h from the low-fidelity density $\tilde{\pi}^{(h)}$, for example, by repeatedly evaluating the log density and its gradients to solve an optimization problem. Second, one must evaluate the unnormalized high-fidelity density $\tilde{\pi}$ to re-weight the N samples drawn from the biasing density μ_h using the estimator (2.12). The first step of learning a biasing density can typically be done in an offline fashion before a specific test function f is provided and incurs a training cost that depends on the fidelity h from evaluating $\tilde{\pi}^{(h)}$. The second step of computing the estimator (2.12) incurs online costs for evaluating the unnormalized high-fidelity density $\tilde{\pi}$ to re-weight the N samples. The combination of both of these steps results in a trade-off between offline versus online computational costs. Initially investing high computational resources to learn a good biasing density will result in higher effective sample sizes for the online phase, thereby reducing the number of expensive high-fidelity evaluations of $\tilde{\pi}$ to achieve a fixed error tolerance. Conversely, investing too little computational resources initially will result in a biasing density with a large χ^2 divergence from the target π and

will suffer from requiring many high-fidelity density evaluations to compensate. This two-step approach for MFIS, as well as other multifidelity methods, combines both the low and high-fidelity models in constrast to raditional model reduction techniques [Quarteroni et al., 2011, Benner et al., 2015] where the low-fidelity model replaces the high-fidelity model. As a result, traditional model reduction provides little guidance on the mathematical formulation of this trade-off and the total combination of online and offline costs.

2.3 Context-aware surrogate models for multifidelity importance sampling

We are concerned with the following optimal trade-off problem for multifidelity importance sampling: Given a tolerance ϵ, such that the mean-squared error of the MFIS estimator (2.12) must be less than or equal to ϵ, determine the optimal surrogate density $\pi^{(h)}$ that minimizes the total computational cost of fitting the biasing density μ_h and computing the estimate (2.12). We refer to such cost-optimal surrogate models as *context-aware* because the fidelity is determined specifically for the online computations of the problem (context) at hand [Peherstorfer, 2019], rather than being prescribed without taking the specific context of multifidelity computations into account as in traditional model reduction [Quarteroni et al., 2011, Benner et al., 2015].

2.3.1 Sub-Gaussian distributions

For importance sampling without a fixed test function f, both estimators (2.3) and (2.4) require that the target density π is absolutely continuous with respect to the biasing density μ meaning that the support of π is contained within the support of μ and that the importance weights (2.2) and (2.5), respectively, have finite variance so that the chi-squared divergence

$\chi^2(\pi \,||\, \mu)$ remains finite. This restriction implies that the tails of the target density π as $\|\boldsymbol{\theta}\| \to \infty$ cannot be much heavier than the tails of the biasing density μ. Sub-Gaussian distributions η are characterized by fast decaying densities, which can be quantified by the Orlicz norm defined for real-valued random variables $X \sim \eta$ as

$$\|X\|_{\psi_2} = \inf\left\{t > 0 \ : \ \mathbb{E}\left[\exp(X^2/t^2)\right] \leq 2\right\} .$$

For vector-valued random variables $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ with multivariate distributions, the Orlicz norm is defined in terms of one dimensional projections

$$\|\boldsymbol{X}\|_{\psi_2} = \sup_{\boldsymbol{v} \in S^{d-1}} \left\|\boldsymbol{v}^\top \boldsymbol{X}\right\|_{\psi_2} ,$$

with $S^{d-1} \subset \mathbb{R}^d$ being the unit sphere: $S^{d-1} = \{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\|_2 = 1\}$. A probability distribution η is said to be sub-Gaussian if any random variable $\boldsymbol{X} \sim \eta$ has $\|\boldsymbol{X}\|_{\psi_2} < \infty$. We refer to the book [Vershynin, 2018, Sec. 2.5, Sec. 3.4] for other equivalent definitions of sub-Gaussian distributions. The class of sub-Gaussian distributions is flexible and includes all Gaussian distributions, distributions with compact support, Gaussian mixtures with finitely many components, and posterior distributions from Bayesian inference where the prior is Gaussian. If $\boldsymbol{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{d\times d})$, where $\mathbf{I}_{d\times d}$ is the d-dimensional identity matrix, then

$$\|\boldsymbol{X}\|_{\psi_2} = \sqrt{\frac{8}{3}}\sigma \,, \tag{2.14}$$

(c.f. Appendix A.1), and is constant factor multiplied by the standard deviation, which controls the rate of decay for the Gaussian density. For distributions η with compact support $\text{supp}(\eta) \subset \mathbb{R}^d$, the random variable $\boldsymbol{X} \in \text{supp}(\eta)$ remains bounded and hence has finite Orlicz norm. In the following Lemma 1, which is a multi-dimensional version [Vershynin, 2018, Proposition 2.5.2 (iv)], we provide a useful characterization of sub-Gaussian distributions

that will be helpful for controlling the chi-squared divergence of the target density π to the biasing density μ. Since Lemma 1 is a technical auxiliary result, we relegate its proof to be found in Appendix A.2.

Lemma 1. *A random vector $\boldsymbol{X} \sim \eta$, and hence the distribution η, is sub-Gaussian if and only if there exists a symmetric positive-definite matrix $\mathbf{A} \succ 0$ such that for all vectors $\boldsymbol{m} \in \mathbb{R}^d$*

$$\mathbb{E}_\eta \left[\exp \left((\boldsymbol{X} - \boldsymbol{m})^\top \mathbf{A} (\boldsymbol{X} - \boldsymbol{m}) \right) \right] < \infty \,. \tag{2.15}$$

Analogous to the Orlicz norm for an isotropic Gaussian (2.14), observe that in the proof of Lemma 1, if $\|\boldsymbol{X}\|_{\psi_2}$ is small then α can be chosen to be large, commensurate with the fast decay of the target density η. Conversely, if $\lambda_{\min}(\mathbf{A})$ is small then $\|\boldsymbol{X}\|_{\psi_2}$ will be large, corresponding to large variance in the example (2.14). We note that any sub-Gaussian distribution such that the negative log density that increases $-\log \eta(\boldsymbol{\theta}) \to \infty$ as $\|\boldsymbol{\theta}\| \to \infty$ must grow at least quadratically fast in order for the expectation (2.15) in Lemma 1 to remain finite.

Remark 2. When η is a Gaussian distribution with covariance $\boldsymbol{\Sigma}$, the matrix $\mathbf{A}$ must be such that $\frac{1}{2}\boldsymbol{\Sigma}^{-1} - \mathbf{A}$ remains positive definite, in which case Lemma 1 is closely related to Fernique's theorem about the tail decay of Gaussian densities. This constraint on $\mathbf{A}$ will translate to a constraint on the biasing density for non-Gaussian target densities as will be made precise in Section 2.3.2.

2.3.2 Bounding the chi-squared divergence

To formulate the optimal trade-off problem described in Section 2.2.3 we need a bound on the mean-squared error, and hence the χ^2 divergence from the high-fidelity target density π to the biasing density μ_h, that depends explicitly on the fidelity h of the surrogate density $\pi^{(h)}$ used

to derive the biasing density. In this section we derive an upper bound on $\chi^2\left(\pi \,||\, \mu_h\right)$ that decomposes into a factor that depends only on the fidelity of the deterministic approximation $\pi^{(h)}$ to π and a second factor that takes into account the quality of the approximation of the learned biasing density μ_h to $\pi^{(h)}$. Such a decomposition is not straightforward for the χ^2 divergence, which is not a strict metric on the space of probability distributions, since it does not admit a triangle inequality. For example, let

$$\pi(x) = a\mathrm{e}^{-ax}, \quad \pi^{(h)}(x) = b\mathrm{e}^{-bx}, \quad \mu_h(x) = c\mathrm{e}^{-cx} \qquad x \geq 0\,,$$

for $a, b, c > 0$, be three exponential distributions. Then

$$\chi^2\left(\pi \,||\, \pi^{(h)}\right) = \int_0^\infty \frac{a^2}{b}\mathrm{e}^{-(2a-b)x}\mathrm{d}x = \frac{a^2}{b(2a-b)}$$

if $a > b/2$ and ∞ otherwise. Taking $a = 2$, $b = 3/2$ and $c = 1$, so that $a < 2b$ and $b < 2c$, but $a \geq 2c$ gives

$$\chi^2\left(\pi \,||\, \pi^{(h)}\right) < \infty, \quad \chi^2\left(\pi^{(h)} \,||\, \mu_h\right) < \infty,$$

but

$$\chi^2\left(\pi \,||\, \pi^{(h)}\right) = \infty\,,$$

meaning that we cannot directly decompose the χ^2 divergence into the product of χ^2 divergences with an intermediate distribution (namely $\pi^{(h)}$). Alternatively, by rewriting the χ^2 divergence as the L^1 norm of the ratio between the target and biasing densities, the Cauchy-Schwarz inequality can be used to give the following decomposition

$$\chi^2\left(\pi \,||\, \mu_h\right) + 1 = \left\|\frac{\pi}{\mu_h}\right\|_{L^1(\pi)} = \left\langle \frac{\pi}{\pi^{(h)}},\ \frac{\pi^{(h)}}{\mu_h} \right\rangle_{L^2(\pi)} \leq \left\|\frac{\pi}{\pi^{(h)}}\right\|_{L^2(\pi)} \left\|\frac{\pi^{(h)}}{\mu_h}\right\|_{L^2(\pi)}. \tag{2.16}$$

An effect of the Cauchy-Schwarz inequality is that the decomposition (2.16) requires the stronger assumption that the density ratios $\pi/\pi^{(h)}$ and $\pi^{(h)}/\mu_h$ are in $L^2(\pi)$ as opposed to comparing the ratio of the target density to the biasing density π/μ_h directly, which only needs to be in $L^1(\pi)$ and is sufficient for the bound on the mean-squared error of the MFIS estimator (2.8) to hold. Although here we restrict the test functions f to be bounded, we could instead allow more general test functions $f \in L^2$ at the cost of placing stronger assumptions on the ratio of densities so that $\left\|\pi/\pi^{(h)}\right\|_{L^\infty(\pi)}, \left\|\pi^{(h)}/\mu_h\right\|_{L^\infty(\pi)} < \infty$ as in [Schillings et al., 2020]. The following four assumptions and theorem are sufficient to decompose the χ^2 divergence as in (2.16) by bounding the L^2 norms of the ratios of densities $\pi/\pi^{(h)}$ and $\pi^{(h)}/\mu_h$.

Assumption 1 (Exponential form of the densities). The densities p, p_h, and q_h have the form

$$\pi(\boldsymbol{\theta}) = \frac{1}{Z}\mathrm{e}^{-\Phi(\boldsymbol{\theta})}, \quad \pi^{(h)}(\boldsymbol{\theta}) = \frac{1}{Z_h}\mathrm{e}^{-\Phi^{(h)}(\boldsymbol{\theta})}, \quad \mu_h(\boldsymbol{\theta}) = \frac{1}{\tilde{Z}_h}\mathrm{e}^{-\tilde{\Phi}^{(h)}(\boldsymbol{\theta})},$$

with potentials (negative log densities) $\Phi, \Phi^{(h)}, \tilde{\Phi}^{(h)} \in \mathcal{C}^2(\Theta)$ that are twice continuously differentiable, normalizing constants $Z, Z_h, \tilde{Z}_h$, and $\Phi^{(h)}(\boldsymbol{\theta}) \to \Phi(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$ as $h \to 0$.

Assumption 2 (Decay of the target density). The target density π is sub-Gaussian with matrix $\mathbf{A}$; see Lemma 1.

Assumption 3 (Error of the surrogate densities). There exists an error function $\delta(h) \geq 0$ and a function $\tau(\boldsymbol{\theta}) \geq 0$, such that

$$\Phi^{(h)}(\boldsymbol{\theta}) \leq \Phi(\boldsymbol{\theta}) + \delta(h)\tau(\boldsymbol{\theta})$$

for all $\boldsymbol{\theta} \in \Theta$, where $\delta(h) \to 0$ as $h \to 0$.

Assumption 4 (Error of biasing densities). There exists a function $\gamma(h) \geq 0$ and a function

$\omega(\boldsymbol{\theta}) \geq 0$ such that for all $h > 0$

$$\tilde{\Phi}^{(h)}(\boldsymbol{\theta}) \leq \Phi^{(h)}(\boldsymbol{\theta}) + \gamma(h)\omega(\boldsymbol{\theta})$$

for all $\boldsymbol{\theta} \in \Theta$.

Remark 3. We borrow the term "potential" from statistical physics to refer to the negative log densities $\Phi, \Phi^{(h)}, \tilde{\Phi}^{(h)}$ since a particle in a potential energy field U at a particular temperature will have the invariant Gibbs distribution e^{-U}.

Assumption 1 is a weak assumption on the densities since any positive density function on Θ may be written in exponential form. For the biasing density μ_h we typically know the normalizing constant $\tilde{Z}_h$ and only write it in exponential form for convenience to simplify the analysis in Theorem 1. The differentiability of the negative log densities in Assumption 1 is only necessary for fitting the Laplace approximation as a biasing density in Section 2.3.3 which requires evaluating the Hessian of the negative log density. Assumption 2 that the target density π is sub-Gaussian is sufficient to avoid heavy tails where the χ^2 divergence can become infinite for Gaussian biasing densities, particularly the Laplace approximation as in Section 2.3.3, resulting in poor importance sampling estimators (2.4). This assumption is also independent of the low-fidelity surrogate densities $\pi^{(h)}$, although since $\pi^{(h)} \to \pi$ pointwise, the surrogate densities will be sub-Gaussian as well, except in pathological cases. The two assumptions 3 and 4 each control one of the terms in the decomposition on the right-hand-side for the bound (2.16). Assumption 3 ensures that the surrogate densities do not decay significantly faster than the high-fidelity target density. Because, $\Phi^{(h)} \to \Phi$ this assumption is straightforward to satisfy by taking h sufficiently small. Assumption 4 places a similar requirement on the learned biasing density that it cannot decay significantly faster than the surrogate density. For the MFIS estimator (2.4) we only need that the biasing density does not decay significantly faster than the high-fidelity target density only, but this

stronger requirement is necessary for the analysis and derivation of a bound on the MSE that depends explicitly on the fidelity h. Note that for both assumptions we assume an asymmetric inequality because the estimator (2.4) may be poor if the tails of the target density are heavier than biasing density, but the reverse with heavier tails for the biasing density will still result in a consistent estimator. We emphasize that Assumptions 2, 3, and 4 are all inherited from constraints of importance sampling as opposed to arising from trading off the surrogate fidelity and costs as discussed in 2.2.3. We also emphasize that Assumption 4 on the growth of the biasing density potential relative to the the surrogate potential Φ does not require that $\gamma(h) \to 0$ as $h \to 0$. Indeed such a requirement is not true in general unless the family of biasing densities contains the true surrogate densities e.g. if the surrogate density itself is Gaussian. Building on the decomposition (2.16), Theorem 1 gives a bound on $\chi^2(\pi \mid\mid \mu_h)$ depending on the fidelity h so long as the biasing density μ_h satisfies the appropriate constraints.

Theorem 1. *Let Assumptions 1, 2, 3, and 4 hold and assume there exist constants $\tau_0, \omega_0 > 0$ such that*

$$\tau(\boldsymbol{\theta}) \leq \|\boldsymbol{\theta}\|^2 + \tau_0, \quad \omega(\boldsymbol{\theta}) \leq \|\boldsymbol{\theta}\|^2 + \omega_0 \,.$$

Let $h_{\max}$ be such that for all $h \leq h_{\max}$

$$\gamma(h) \leq \frac{1}{4}\lambda_{\min}(\mathbf{A}) \,, \tag{2.17}$$

with $\mathbf{A}$ being the matrix from Assumption 2 and $\lambda_{\min}(\mathbf{A})$ being its smallest eigenvalue, then for all h sufficiently small we have that

$$\chi^2(\pi \mid\mid \mu_h) + 1 \leq K_0 \mathrm{e}^{K_1\delta(h) + K_2\gamma(h)} \tag{2.18}$$

where K_0, K_1, K_2 are all positive constants independent of h.

Proof of Theorem 1. By Assumption 2, π is sub-Gaussian with matrix $\mathbf{A} \succ 0$ so that by Lemma 1

$$\frac{1}{Z}\int_\Theta \exp\left(\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta} - \Phi(\boldsymbol{\theta})\right) \mathrm{d}\boldsymbol{\theta} < \infty\,.$$

Recall that Z is the normalizing constant from Assumption 1.

Part 1: Bounding high-fidelity to surrogate ratio

The first term on the right-hand-side of Equation (2.16) can be bounded using Assumption 3:

$$\begin{aligned}\left\|\frac{\pi}{\pi^{(h)}}\right\|^2_{L^2(\pi)} &= \frac{1}{Z}\left(\frac{Z_h}{Z}\right)^2 \int_\Theta \exp\left\{2\left(\Phi^{(h)}(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})\right) - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta} \\ &\leq \frac{1}{Z}\left(\frac{Z_h}{Z}\right)^2 \int_\Theta \exp\left\{2\delta(h)\left(\|\boldsymbol{\theta}\|^2 + \tau_0\right) - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta}\,.\end{aligned}$$

Re-writing this last line gives

$$\left\|\frac{\pi}{\pi^{(h)}}\right\|^2_{L^2(\pi)} \leq \frac{1}{Z}\left(\frac{Z_h}{Z}\right)^2 \exp\left(2\tau_0\delta(h)\right) \int_\Theta \exp\left\{2\delta(h)\|\boldsymbol{\theta}\|^2 - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta}\,. \tag{2.19}$$

Now the two dependencies of the right-hand side of (2.19) on the fidelity h are through the ratio Z_h/Z and through $\delta(h)$. For now we just bound the integral on the right-hand side of (2.19), which is finite since $\mathbf{A} \succ 2\delta(h)\mathbf{I}$ for all h sufficiently small. Adding and subtracting $\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta}$ in (2.19) gives

$$\begin{aligned}\left\|\frac{\pi}{\pi^{(h)}}\right\|^2_{L^2(\pi)} &\leq \frac{1}{Z}\left(\frac{Z_h}{Z}\right)^2 \exp\left(2\tau_0\delta(h)\right) \int_\Theta \exp\left\{2\delta(h)\|\boldsymbol{\theta}\|^2 - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{Z}\left(\frac{Z_h}{Z}\right)^2 \exp\left(2\tau_0\delta(h)\right) \int_\Theta \exp\left\{-\boldsymbol{\theta}^\top\left(\mathbf{A} - 2\delta(h)\mathbf{I}\right)\boldsymbol{\theta} + \boldsymbol{\theta}^\top\mathbf{A}\boldsymbol{\theta} - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta}\,.\end{aligned}$$

Putting this together with the fact that $\mathbf{A} - 2\delta(h)\mathbf{I} \succ 0$ gives

$$\left\|\frac{\pi}{\pi^{(h)}}\right\|^2_{L^2(\pi)} \leq \frac{1}{Z}\left(\frac{Z_h}{Z}\right)^2 \exp\left(2\tau_0\delta(h)\right) \int_\Theta \exp\left\{\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta} - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta} \tag{2.20}$$

to complete the bound of the first term on the right-hand side of Equation (2.16).

Part 2: Bounding surrogate to biasing density ratio

The second term on the right-hand side of Equation (2.16) is bounded in a similar fashion. By Assumption 4 we can bound

$$\begin{aligned}\left\|\frac{\pi^{(h)}}{\mu_h}\right\|_{L^2(\pi)}^2 &= \frac{1}{Z}\left(\frac{\tilde{Z}_h}{Z_h}\right)^2 \int_\Theta \exp\left\{2\left(\tilde{\Phi}^{(h)}(\boldsymbol{\theta}) - \Phi^{(h)}(\boldsymbol{\theta})\right) - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta} \\ &\leq \frac{1}{Z}\left(\frac{\tilde{Z}_h}{Z_h}\right)^2 \int_\Theta \exp\left\{2\gamma(h)\left(\|\boldsymbol{\theta}\|^2 + \omega_0\right) - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta} \\ &= \frac{1}{Z}\left(\frac{\tilde{Z}_h}{Z_h}\right)^2 \exp\left(2\omega_0\gamma(h)\right) \int_\Theta \exp\left\{2\gamma(h)\|\boldsymbol{\theta}\|^2 - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta}\,.\end{aligned}$$

Again we add and subtract $\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta}$ to obtain

$$\left\|\frac{\pi^{(h)}}{\mu_h}\right\|_{L^2(\pi)}^2 \leq \frac{1}{Z}\left(\frac{\tilde{Z}_h}{Z_h}\right)^2 \exp\left(2\omega_0\gamma(h)\right) \int_\Theta \exp\left\{-\boldsymbol{\theta}^\top\left(\mathbf{A} - 2\gamma(h)\mathbf{I}\right)\boldsymbol{\theta} + \boldsymbol{\theta}^\top\mathbf{A}\boldsymbol{\theta} - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta}\,.$$

Using this with the fact that $\mathbf{A} - 2\gamma(h)\mathbf{I} \succeq 0$ for all $h \leq h_{\max}$ gives

$$\left\|\frac{\pi^{(h)}}{\mu_h}\right\|_{L^2(\pi)}^2 \leq \frac{1}{Z}\left(\frac{\tilde{Z}_h}{Z_h}\right)^2 \exp\left(2\omega_0\gamma(h)\right) \int_\Theta \exp\left\{\boldsymbol{\theta}^\top\mathbf{A}\boldsymbol{\theta} - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta}\,. \tag{2.21}$$

Multiplying the right-hand sides of the bounds (2.20) and (2.21) and then taking the square root gives together with (2.16) that

$$\left\|\frac{\pi}{\mu_h}\right\|_{L^1(\pi)} \leq \frac{1}{Z}\left(\frac{\tilde{Z}_h}{Z}\right) \exp\left\{\delta(h)\tau_0 + \gamma(h)\omega_0\right\} \int_\Theta \exp\left\{\boldsymbol{\theta}^\top\mathbf{A}\boldsymbol{\theta} - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta} \tag{2.22}$$

holds. The integral is independent of h, so it remains to bound the ratio of normalizing constants.

Part 3: Bounding ratio of normalizing constants

In general, if $\pi^{(h)}$ is not in the family of biasing densities then we may have $\tilde{Z}_h \neq Z_h$, and thus,

$$\frac{\tilde{Z}_h}{Z} \not\to 1$$

as $h \to 0$. Instead we just give a constant upper bound on $\tilde{Z}_h$ that is independent of the fidelity h. By Assumption 1, the normalizing constant $\tilde{Z}_h$ satisfies

$$\begin{aligned}
\tilde{Z}_h &= \int_\Theta \exp\left\{-\tilde{\Phi}^{(h)}(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta} \\
&= \int_\Theta \exp\left\{-\tilde{\Phi}^{(h)}(\boldsymbol{\theta}) + \Phi^{(h)}(\boldsymbol{\theta}) - \Phi^{(h)}(\boldsymbol{\theta}) + \Phi(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})\right\} \mathrm{d}\boldsymbol{\theta} \\
&= Z \int_\Theta \exp\left\{-\tilde{\Phi}^{(h)}(\boldsymbol{\theta}) + \Phi^{(h)}(\boldsymbol{\theta}) - \Phi^{(h)}(\boldsymbol{\theta}) + \Phi(\boldsymbol{\theta})\right\}\ p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\,.
\end{aligned}$$

Dividing by Z and using Assumptions 3 and 4 we have

$$\frac{\tilde{Z}_h}{Z} \leq \int_\Theta \exp\left\{-\delta(h)(\|\boldsymbol{\theta}\|^2 + \tau_0) - \gamma(h)(\|\boldsymbol{\theta}\|^2 + \omega_0)\right\} \pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \leq 1\,, \tag{2.23}$$

because the term inside the exponential is less than or equal to 0 and π is a density. Finally, combining the bounds (2.20), (2.21), and (2.23) gives the result

$$\chi^2\left(\pi \mid\mid \mu_h\right) + 1 = \left\| \frac{\pi}{\mu_h} \right\|_{L^1(\pi)} \leq \exp\left\{\delta(h)\tau_0 + \gamma(h)\omega_0\right\} \mathbb{E}_\pi\left[\exp\left(\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta}\right)\right],$$

where the expectation is independent of h. Here

$$K_0 = \mathbb{E}_\pi\left[\exp\left(\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta}\right)\right], \quad K_1 = \tau_0, \quad K_2 = \omega_0$$

are all independent of the fidelity h. □

The requirement in Theorem 1 that $\gamma(h) \leq \lambda_{\min}(\mathbf{A})/4$, allows us to obtain an alternative

bound to (2.18) which depends only on the error of the surrogate potentials.

$$\chi^2\left(\pi \,||\, \mu_h\right) + 1 \leq \tilde{K}_0 \mathrm{e}^{K_1 \delta(h)}\,, \tag{2.24}$$

where the constant $\tilde{K}_0$ now absorbs the dependency on the approximation μ_h and is given by

$$\tilde{K}_0 = K_0 \mathrm{e}^{K_2 \lambda_{\min}(\mathbf{A})/4} \geq K_0 \mathrm{e}^{K_2 \gamma(h)}\,. \tag{2.25}$$

Subtracting 1 from both sides of the bound (2.24) and taking the limit $h \to 0$ gives

$$\lim_{h \to 0} \chi^2\left(\pi \,||\, \mu_h\right) \leq \tilde{K}_0 - 1\,,$$

with $\tilde{K}_0 \geq K_0 > 1$, which is determined by the choice of biasing densities μ_h. Note that when π and μ_h are both Gaussian, as is the case for linear Bayesian inverse problems discussed in Section 1.3.2, the exact χ^2 divergence also has an exponential form similar to (2.24), c.f. (B.6) in Appendix B.

Remark 4. The assumption that $\tau(\boldsymbol{\theta}) \leq \|\boldsymbol{\theta}\|^2 + \tau_0$ holds is similar to the pointwise Assumption 4.8 in [Stuart, 2010, Theorem 4.6]. In [Stuart, 2010], the pointwise bound can grow faster with respect to $\boldsymbol{\theta}$ than in our case because there the Hellinger distance (3.8), which is upper-bounded by the χ^2 divergence, is considered.

2.3.3 Laplace approximation as biasing density

In this section we consider the explicit choice of biasing density μ_h to be a Laplace approximation of the surrogate density $\pi^{(h)}$. A Laplace approximation to the surrogate density $\pi^{(h)}$ is a Gaussian approximation whose mean is the mode

$$\boldsymbol{\mu}_h^{\mathrm{LAP}} = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} \; -\log \tilde{\pi}^{(h)}(\boldsymbol{\theta}) = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} \Phi^{(h)}(\boldsymbol{\theta}), \tag{2.26}$$

and whose covariance is the negative inverse Hessian of the density evaluated at the mode

$$\boldsymbol{\Sigma}_h^{\mathrm{LAP}} = -\left[\nabla\nabla^\top \log \tilde{\pi}^{(h)}\left(\boldsymbol{\mu}_h^{\mathrm{LAP}}\right)\right]^{-1} = \left[\nabla\nabla^\top \Phi^{(h)}\left(\boldsymbol{\mu}_h^{\mathrm{LAP}}\right)\right]^{-1} . \tag{2.27}$$

Importantly, the Laplace approximation can be computed by evaluating only the unnormalized density $\tilde{\pi}^{(h)}$ and its derivatives. Note that in the case of a multimodal surrogate density with two or more global optima for the potential $\Phi^{(h)}$, we may choose any of the modes; see Figure 2.1. Furthermore, determining the global optima for (2.26) in practice may be infeasible so the Laplace mean is taken to be a local mode instead. Multimodal distributions can be problematic if the biasing density decays too quickly to place sufficient probability in the regions of the other modes, violating Assumption 4, and therefore causing the MFIS estimator (2.12) to have large or even infinite variance. Another potential issue is that a Laplace approximation may not exist for certain distributions where either the covariance matrix $\boldsymbol{\Sigma}_h^{\mathrm{LAP}}$ or Hessian at the mode is not full-rank. This is the case, for example, in a uniform distribution where the Hessian of the log density is zero everywhere. The work [Schillings et al., 2020] provides more in-depth discussion about Laplace approximations as biasing densities when the covariance matrix becomes singular, which is often the case in concentrated posteriors, c.f. Section 1.3.2.

The following proposition gives conditions on the surrogate densities $\pi^{(h)}$ for the Laplace approximation to exist and shows that a $\gamma(h)$ to satisfy Assumption 4 exists. However, the $\gamma(h)$ provided by Proposition 1 may not automatically satisfy the additional assumption that $\gamma(h) \leq \lambda_{\min}(\mathbf{A})$ in (2.17) of Theorem 1, which may still need to verified independently, potentially with an alternative $\gamma(h)$.

Proposition 1. *Let Assumption 1 hold and assume there exists a $\sigma^2_{\min} > 0$, independent of h, such that*

$$\boldsymbol{\theta}^\top \boldsymbol{\Sigma}_h^{\mathrm{LAP}} \boldsymbol{\theta} \geq \sigma^2_{\min} \|\boldsymbol{\theta}\|^2 , \tag{2.28}$$

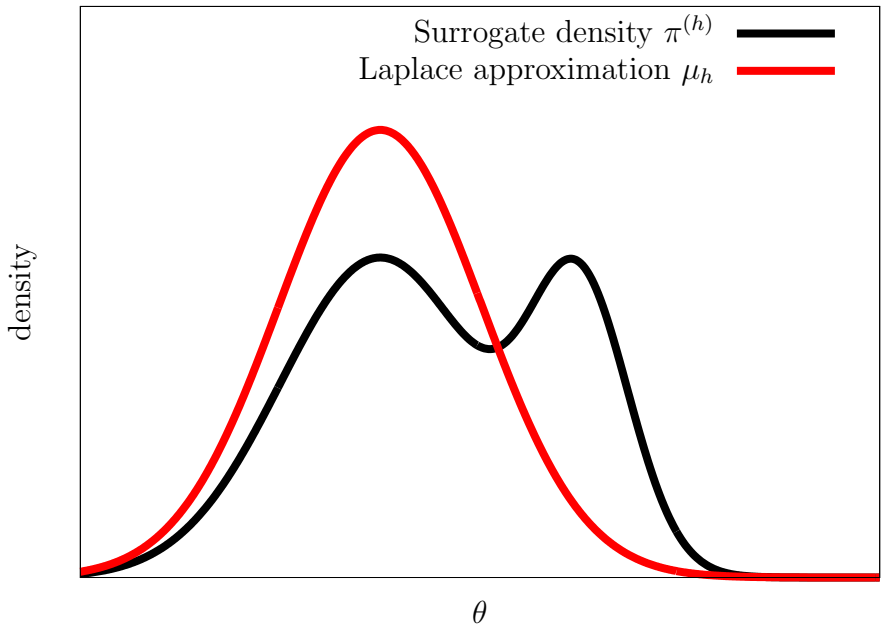


Figure 2.1: A Laplace approximation for the density of a Gaussian mixture with two modes.

for all $\boldsymbol{\theta} \in \Theta$*. Further, assume there exist constants* $V \in \mathbb{R}$ *and* $v > 0$ *such that*

$$\Phi^{(h)}(\boldsymbol{\theta}) \geq V - v \left\|\boldsymbol{\theta}\right\|^2 \tag{2.29}$$

for all h*. Finally, let* $B_R = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\| \leq R\}$ *be the ball of radius* R *centered at* 0*, and assume that for all* $D > 0$*, there exists an* $R(D) > 0$ *such that for all* $\boldsymbol{\theta} \notin B_{R(D)}$ *and all* $h > 0$

$$\Phi^{(h)}(\boldsymbol{\theta}) \geq D\,. \tag{2.30}$$

Then, the Laplace approximation satisfies Assumption 4 for all h *sufficiently small.*

Proof. By Assumption 1, a Laplace approximation

$$\tilde{\Phi}^{(h)}(\boldsymbol{\theta}) = \Phi^{(h)}\left(\boldsymbol{\mu}_h^{\mathrm{LAP}}\right) + \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\mu}_h^{\mathrm{LAP}}\right)^\top \left[\nabla\nabla^\top \Phi^{(h)}(\boldsymbol{\mu}_h^{\mathrm{LAP}})\right]^{-1}\left(\boldsymbol{\theta} - \boldsymbol{\mu}_h^{\mathrm{LAP}}\right)$$

is the second-order Taylor expansion of $\Phi^{(h)}$ around one of the modes $\boldsymbol{\mu}_h^{\mathrm{LAP}}$.

The first derivative is zero since it is expanded around a minimizer. Therefore,

$$\tilde{\Phi}^{(h)}(\boldsymbol{\theta}) - \Phi^{(h)}(\boldsymbol{\theta}) = -R_h(\boldsymbol{\theta})\,,$$

where $R_h(\boldsymbol{\theta})$ is the remainder of higher order terms from the Taylor expansion. The bound (2.28) implies that

$$\boldsymbol{\theta}^\top \left(\boldsymbol{\Sigma}_h^{\mathrm{LAP}}\right)^{-1} \boldsymbol{\theta} \leq \frac{1}{\sigma_{\min}^2} \left\|\boldsymbol{\theta}\right\|^2 ,$$

and when combined with the bound (2.29) gives

$$\begin{aligned}\tilde{\Phi}^{(h)}(\boldsymbol{\theta}) - \Phi^{(h)}(\boldsymbol{\theta}) &\leq \tilde{\Phi}^{(h)}(\boldsymbol{\theta}) - V + v\left\|\boldsymbol{\theta}\right\|^2 \\ &\leq \Phi^{(h)}\left(\boldsymbol{\mu}_h^{\mathrm{LAP}}\right) + \frac{1}{2\sigma_{\min}^2}\left\|\boldsymbol{\theta} - \boldsymbol{\mu}_h^{\mathrm{LAP}}\right\|^2 - V + v\left\|\boldsymbol{\theta}\right\|^2 .\end{aligned}$$

Combining this with the fact that $\left\|\boldsymbol{v} - \boldsymbol{w}\right\|^2 \leq 2\left\|\boldsymbol{v}\right\|^2 + 2\left\|\boldsymbol{w}\right\|^2$ yields

$$\tilde{\Phi}^{(h)}(\boldsymbol{\theta}) - \Phi^{(h)}(\boldsymbol{\theta}) \leq \Phi^{(h)}\left(\boldsymbol{\mu}_h^{\mathrm{LAP}}\right) + \left(\frac{1}{\sigma_{\min}^2} + v\right)\left\|\boldsymbol{\theta}\right\|^2 + \frac{1}{\sigma_{\min}^2}\left\|\boldsymbol{\mu}_h^{\mathrm{LAP}}\right\|^2 - V\,.$$

Now we claim that the terms $\Phi^{(h)}(\boldsymbol{\mu}_h^{\mathrm{LAP}})$ and $\left\|\boldsymbol{\mu}_h^{\mathrm{LAP}}\right\|^2$ can be bounded independent of h. Let $D = \Phi(0) + 1$ and consider that, by assumption, there exists a ball $B_{R(D)}$ such that

$$\Phi^{(h)}(\boldsymbol{\theta}) \geq \Phi(0) + 1\,, \quad \forall \boldsymbol{\theta} \notin B_{R(D)}\,.$$

By Assumption 1, we know that $\Phi_h(0) \to \Phi(0)$ and so that for all h sufficiently small, there exist points $\boldsymbol{\theta}_h'$, such that $\Phi_h(\boldsymbol{\theta}_h') \leq \Phi(0) + 1$. Hence, the minimizers $\boldsymbol{\mu}_h^{\mathrm{LAP}} \in B_R$ for all h sufficiently small. Thus, there are constants $B_1, B_2 > 0$ independent of h such that

$\Phi^{(h)}\left(\boldsymbol{\mu}_h^{\mathrm{LAP}}\right) \leq B_1$ and $\left\|\boldsymbol{\mu}_h^{\mathrm{LAP}}\right\|^2 \leq B_2$. Thus, by setting

$$\gamma(h) = \frac{1}{\sigma_{\min}^2} + v, \quad \omega(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^2 + \omega_0, \quad \omega_0 = \frac{B_1 + B_2/\sigma_{\min}^2 - V}{\sigma_{\min}^{-2} + v}$$

Assumption 4 holds. □

The assumption (2.28) in Proposition 1 ensures existence of the Laplace approximation for each $h > 0$ and moreover ensures that as $h \to 0$ they remain non-singular. This instead can be viewed as a constraint for the Hessians of the surrogate potentials $\Phi^{(h)}$ at the mode to remain bounded as $h \to 0$:

$$\alpha_{\max}\mathbf{I} - \nabla\nabla^\top \Phi^{(h)}(\boldsymbol{\mu}_h^{\mathrm{LAP}}) \succ 0\,,$$

for some $\alpha_{\max} > 0$. The condition (2.30) implies that $\Phi^{(h)}(\boldsymbol{\theta}) \to \infty$ as $\|\boldsymbol{\theta}\| \to \infty$ uniformly in h ensuring that a global minimizer (2.26) exists for each potential $\Phi^{(h)}$ as well as preventing the sequence of global minimizers being unbounded as $h \to 0$. We recall that although a global minimizer exists for each potential $\Phi^{(h)}$, it does not necessarily have to be unique. Finally, the last condition (2.29) is analogous to Assumption 2.6(i) from [Stuart, 2010] and is necessary to satisfy Assumption 4 needed for Theorem 1.

2.3.4 Trading off fidelity and costs of surrogate model for MFIS

In this section we consider the trade-off problem to select the optimal fidelity h for learning the Laplace approximation as a biasing density and computing the MFIS estimator (2.12).

Offline and online costs of MFIS with Laplace approximation as biasing density

The dominant costs of estimating the quantity of interest (2.1) with the MFIS estimator (2.12) are the offline training costs to learn the biasing density using the unnormalized

surrogate density $\tilde{\pi}^{(h)}$ and the online costs to evaluate the importance weights $\tilde{w}^{(h)}(\boldsymbol{\theta}^{[i]})$ for $i = 1, \dots, N$ in (2.13). In the following analysis we assume that the costs of drawing samples from the biasing density $\boldsymbol{\theta}^{[i]} \sim \mu_h$ as well as the costs of evaluating the biasing density $\mu_h(\boldsymbol{\theta}^{[i]})$ are negligible compared to the costs C^{high} and $c(h)$ of evaluating the high-fidelity and low-fidelity surrogate densities, respectively. Although here we consider the Laplace approximation as the biasing density due to its favorable approximation guarantees and ease to compute, the analysis in this section apply to a general biasing density for which Theorem 1 applies and for which the costs to evaluate the density and sample from are negligible. For the offline phase for learning the biasing density we assume that we perform N_0 evaluations of the unnormalized surrogate density $\tilde{\pi}^{(h)}$ to fit the Laplace approximation μ_h. Therefore, the total costs of the offline phase are modeled as

$$c_{\text{offline}} = N_0 c(h)\,. \tag{2.31}$$

For example, in the numerical examples in Section 2.5, N_0 is the number of density evaluations used in Newton's method to achieve machine precision in the gradient of $-\log \tilde{\pi}^{(h)}$ and evaluate the Hessian at the mode. After learning the Laplace approximation the estimator (2.12) can be computed in an online phase, which involves evaluating the high-fidelity target density at N samples, whose total costs can be modeled as

$$c_{\text{online}} = N C^{\text{high}}\,. \tag{2.32}$$

Combining both online and offline costs, we model the total cost of the MFIS estimator (2.12) as

$$\text{cost}(\hat{f}_{h,N}^{\text{MFIS}}) = c_{\text{online}} + c_{\text{offline}} = N C^{\text{high}} + N_0 c(h)\,. \tag{2.33}$$

Although the online costs are independent of the surrogate density $\pi^{(h)}$ used to fit the Laplace approximation, as we showed in the bound (2.8) and Theorem 1, the mean-squared error of the estimator (2.12) is controlled by the effective sample size (2.9) and, in turn, the fidelity h.

Context-aware importance sampling

In this section we derive the context-aware importance sampling (CAIS) estimator based on an optimization problem for the trade-off described in Section 2.2.3. Consider the following optimization problem for the optimal fidelity h and number of online samples N that minimizes the total cost (2.33) of the MFIS estimator given the constraint that its mean-squared error is bounded above by a tolerance ϵ:

$$\begin{aligned} \min_{h>0,\ N\in\mathbb{N}} \quad & NC^{\text{high}} + N_0 c(h)\,, \\ \text{such that} \quad & \frac{4\tilde{K}_0}{N}\mathrm{e}^{K_1\delta(h)} \leq \epsilon\,, \end{aligned} \tag{2.34}$$

where the constants $\tilde{K}_0, K_1$ are given by Theorem 1 and (2.25) and the function $\delta(h)$ is as in Assumption 3. Using the bounds (2.8) and (2.24) derived from Theorem 1, any solution (h, N) to (2.34) is guaranteed to provide an estimator $\hat{f}^{\text{MFIS}}_{h,N}$ with mean-squared error bounded above by ϵ. Because the optimization problem (2.34) is over integer sample sizes, we instead solve a relaxed optimization problem where we allow the arguments to be real-valued. Lemma 2 below defines this relaxation of the optimization problem and shows that a unique solution exists under mild conditions.

Lemma 2. *Let $c(\hat{h})$ and $e(\hat{h})$ be continuous non-negative convex functions, at least one of which is strictly convex. Let further $c(\hat{h})$ be monotonically decreasing and $e(\hat{h})$ be monotonically increasing as $\hat{h} \to \infty$. Let $\epsilon > 0$ be a tolerance and $N_0 \in \mathbb{N}$ and C^{high} be constants*

independent of $\hat{h}$*. Then, there exists a unique solution* $(\hat{h}^*, \hat{N}^*) \in (0,\infty) \times (0,\infty)$ *of*

$$\begin{aligned} \min_{\hat{h}>0,\ \hat{N}>0} \quad & \hat{N} C^{\text{high}} + N_0 c(\hat{h})\,, \\ \text{such that} \quad & \frac{1}{\hat{N}} e(\hat{h}) \leq \epsilon\,. \end{aligned} \tag{2.35}$$

Proof of Lemma 2. We proceed as follows: first we show that if a solution exists it cannot occur at zero or infinity (i.e. too high or low fidelity), then we show that a solution exists over a compact interval, and finally show its uniqueness. For any $\hat{h}$, the optimal $\hat{N}$ is the one that achieves equality in the constraint

$$\hat{N} = \frac{e(\hat{h})}{\epsilon}\,. \tag{2.36}$$

Plugging this into the objective function gives the minimization problem over $\hat{h}$ only.

$$\min_{\hat{h}>0} \quad C^{\text{high}} \frac{e(\hat{h})}{\epsilon} + N_0 c(\hat{h})\,. \tag{2.37}$$

We first show that the infimum of the objective function cannot occur as $\hat{h} \to \infty$ or as $\hat{h} \to 0$. Since $c(\hat{h})$ is non-negative and decreasing as $h \to \infty$ we know that $c(\hat{h}) \to c_0$ for some constant $c_0 \geq 0$. Moreover, $e(\hat{h})$ is increasing, so we know that there exists an $\hat{h}_{\max} < \infty$, such that any optimal solution $\hat{h}^*$ must satisfy $\hat{h}^* \leq \hat{h}_{\max}$. Similarly, since $e(\hat{h})$ is non-negative and decreasing as $\hat{h} \to 0$ we know that $e(\hat{h}) \to e_0$ for some constant $e_0 \geq 0$ as $\hat{h} \to 0$. Moreover, $c(\hat{h})$ is increasing as $\hat{h} \to 0$, and since the objective function (2.37) is monotonically increasing as $\hat{h} \to 0$, we know that there exists an $\hat{h}_{\min} > 0$, such that any optimal solution $\hat{h}^*$ must satisfy $\hat{h}^* \geq \hat{h}_{\min}$. Hence

$$\min_{\hat{h}>0} \quad C^{\text{high}} \frac{e(\hat{h})}{\epsilon} + N_0 c(\hat{h}) = \min_{\hat{h} \in [\hat{h}_{\min}, \hat{h}_{\max}]} \quad C^{\text{high}} \frac{e(\hat{h})}{\epsilon} + N_0 c(\hat{h})$$

Since the objective function is continuous over a compact set, we know that a minimizer exists. Finally, the sum of a strictly convex function and a convex function is strictly convex, so we know that this objective function is strictly convex in $\hat{h}$, and therefore the minimizer is unique. □

Remark 5. We use the notation $\hat{h}$ and $\hat{N}$ in the relaxed optimization problem (2.35) to indicate that the variable $\hat{N}$ may not be an integer and that the optimal solution $(\hat{h}^*, \hat{N}^*)$ of (2.35) may not necessarily be the same as the optimal solution of (2.34).

To apply Lemma 2, let δ from Assumption 3 also be convex, although not necessarily strictly convex, and set

$$e(\hat{h}) = 4\tilde{K}_0 \mathrm{e}^{K_1 \delta(h)} \,. \tag{2.38}$$

The function $e(\hat{h})$ is non-negative and strictly convex since it is the composition of the increasing and strictly convex function $x \mapsto \mathrm{e}^x$ and the convex function δ. Let $(\hat{h}^*, \hat{N}^*)$ be the solution to the optimization problem (2.35) with $e(\hat{h})$ given by (2.38). We define the context-aware importance sampling (CAIS) estimator

$$\hat{f}^{\mathrm{CAIS}}_{h^*,N^*} = \hat{f}^{\mathrm{MFIS}}_{h^*,N^*} \,, \tag{2.39}$$

where $h^* = \hat{h}^*$ and $N^* = \left\lceil \hat{N}^* \right\rceil$. We refer to the surrogate model or density $\pi^{(h^*)}$ as a context-aware surrogate model precisely because it takes the cost of the online importance sampling into consideration during the optimization (2.34), c.f. Section 1.1.1.

Because $N^* \geq \hat{N}^*$ the CAIS estimator (2.39) is guaranteed to have a mean-squared error bounded by ϵ. Moreover, in practical applications only a subset of fidelities $h_1 > \ldots > h_L$ may be available. In this case, assuming at least $h_L \leq \hat{h}^*$, we choose the largest (poorest)

fidelity that does not exceed $\hat{h}^*$,

$$h^* = \max\left\{\ell \ : \ h_\ell \leq \hat{h}^*\right\},$$

which guarantees that the mean-squared error of (2.39) will be bounded above by ϵ.

Cost complexity bounds of CAIS

In this section we derive an upper bound on the cost complexity of the CAIS estimator (2.39) in terms of the MSE tolerance ϵ and compare with the cost complexity of (2.12) at a fixed fidelity $\bar{h}$ when the biasing density is given by the Laplace approximation.

Theorem 2. *Suppose that both Theorem 1 and Proposition 1 apply. Consider a tolerance* $0 < \epsilon \leq 1$ *and set* $K_0' = 4\,\|f\|_{L^\infty}^2\,\tilde{K}_0 + 1$, *where* $\tilde{K}_0$ *is the constant in Equation (2.25).*

1. *If the surrogate density evaluation costs grow as* $c(h) = \beta^{1/h}$ *with the fidelity* h *and the surrogate error decays as* $\delta(h) = \alpha^{-1/h}$ *in Assumption 3, with* $\alpha, \beta > 1$, *and we restrict* $h \in (0, \log(\alpha)/2]$, *then the total costs (2.33) of the context-aware importance sampling estimator (2.39) are bounded as*

$$\mathrm{cost}(\hat{f}^{\mathrm{CAIS}}_{h^*,N^*}) \leq \overline{\mathrm{cost}}(\hat{f}^{\mathrm{CAIS}}_{h^*,N^*}) = \frac{C^{\mathrm{high}} K_0'}{\epsilon} \mathrm{e}^{K_1 \epsilon^{1/(1+\log_\alpha \beta)}} + N_0 \epsilon^{-1/(1+\log_\beta \alpha)}\,. \tag{2.40}$$

2. *If instead* $c(h) = h^{-\beta}$ *and* $\delta(h) = h^\alpha$ *with* $\alpha, \beta > 0$, *then the costs are bounded as*

$$\mathrm{cost}(\hat{f}^{\mathrm{CAIS}}_{h^*,N^*}) \leq \overline{\mathrm{cost}}(\hat{f}^{\mathrm{CAIS}}_{h^*,N^*}) = \frac{C^{\mathrm{high}} K_0'}{\epsilon} \mathrm{e}^{K_1 \epsilon^{\alpha/(\alpha+\beta)}} + N_0 \epsilon^{-\beta/(\alpha+\beta)}\,. \tag{2.41}$$

Proof of Theorem 2. Recall that the CAIS estimator (2.39) solves the relaxed optimization

problem (2.35) with

$$e(h) = 4 \left\| f \right\|_{L^\infty}^2 \tilde{K}_0 \mathrm{e}^{K_1 \delta(h)} .$$

The function $\delta(h) = \alpha^{-1/h}$ is convex over the interval $(0, \log(\alpha)/2]$ and decreasing as $h \to 0$, while $c(h) = \beta^{1/h}$ is convex and increasing as $h \to 0$, so the assumptions of Lemma 2 are valid and there exists a unique solution $\hat{h}^* > 0$ and $\hat{N}^* > 0$.

We can remove the constraint in (2.35) to instead minimize

$$\min_{\hat{h}>0} \quad \frac{4 \left\| f \right\|_{L^\infty}^2 \tilde{K}_0 C^{\text{high}}}{\epsilon} \mathrm{e}^{K_1 \delta(\hat{h})} + N_0 c(\hat{h}), \tag{2.42}$$

as is done in (2.37). By setting the derivative of (2.42) with respect to $\hat{h}$ to zero, the optimal solution $\hat{h}^*$ satisfies

$$\frac{4 \left\| f \right\|_{L^\infty}^2 \tilde{K}_0 K_1 C^{\text{high}} \log \alpha}{N_0 \log \beta} \mathrm{e}^{K_1 \alpha^{-1/\hat{h}^*}} = \epsilon (\alpha\beta)^{1/\hat{h}^*} . \tag{2.43}$$

Since $\hat{h}^* > 0$, the left-hand-side of (2.43) is bounded below by a constant independent of ϵ and therefore $\hat{h}^* \to 0$ as $\epsilon \to 0$ to balance the right-hand-side. In particular, we must have $1/\hat{h}^* \in \mathcal{O}(\log_{\alpha\beta} \epsilon^{-1})$ as $\epsilon \to 0$. This motivates setting $h^\dagger = 1/\log_{\alpha\beta} \epsilon^{-1}$, which scales as the same rate as the optimal solution $\hat{h}^*$. By plugging in $h^\dagger$, the corresponding number of samples needed to achieve the constraint of ϵ will be

$$N^\dagger = \left\lceil \frac{4 \left\| f \right\|_{L^\infty}^2 \tilde{K}_0}{\epsilon} \mathrm{e}^{K_1 \epsilon^{1/(1+\log_\alpha \beta)}} \right\rceil \leq \frac{4 \left\| f \right\|_{L^\infty}^2 \tilde{K}_0}{\epsilon} \mathrm{e}^{K_1 \epsilon^{1/(1+\log_\alpha \beta)}} + 1 ,$$

where we have used that $\log_{\alpha\beta} \epsilon = \frac{\log_\alpha \epsilon}{1+\log_\alpha \beta} = \frac{\log_\beta \epsilon}{1+\log_\beta \alpha}$. Since $\epsilon \leq 1$ we know that $\mathrm{e}^{K_1 \epsilon^{1/(1+\log_\alpha \beta)}}/\epsilon > 1$, and so

$$N^\dagger \leq K_0' \frac{\mathrm{e}^{K_1 \epsilon^{1/(1+\log_\alpha \beta)}}}{\epsilon} .$$

Because the CAIS estimator (2.39) uses the optimal parameters $h^* > 0$ and $N^* \in \mathbb{N}$ that solve (2.35) we know that plugging $(h^\dagger, N^\dagger)$ into the formula for the total costs (2.33) will give an upper bound, which we denote by $\overline{\text{cost}}$, on the total computational costs of the CAIS estimator

$$\text{cost}(\hat{f}^{\text{CAIS}}_{h^*,N^*}) \leq \frac{C^{\text{high}} K_0'}{\epsilon} \mathrm{e}^{K_1 \epsilon^{1/(1+\log_\alpha \beta)}} + N_0 \epsilon^{-1/(1+\log_\beta \alpha)} .$$

Now consider $c(h) = h^{-\beta}$ and $\delta(h) = h^\alpha$ with $\alpha, \beta \geq 1$, which again satisfy the necessary assumptions of Lemma 2. As in the previous case, set the derivative of (2.37) to zero to find that the optimal solution satisfies

$$\frac{4 \left\| f \right\|^2_{L^\infty} C^{\text{high}} \tilde{K}_0 K_1}{N_0} \left(\frac{\alpha}{\beta} \right) \mathrm{e}^{K_1 \hat{h}^\alpha} \hat{h}^{\alpha+\beta} = \epsilon ,$$

so that, by the same reasoning as before, $\hat{h}^* \in \mathcal{O}(\epsilon^{1/(\alpha+\beta)})$ as $\epsilon \to 0$. If we now set $h^\dagger = \epsilon^{1/(\alpha+\beta)}$, then the number of samples needed is

$$N^\dagger = \left\lceil \frac{4 \left\| f \right\|^2_{L^\infty} \tilde{K}_0}{\epsilon} \mathrm{e}^{K_1 \epsilon^{\alpha/(\alpha+\beta)}} \right\rceil \leq \frac{4 \left\| f \right\|^2_{L^\infty} \tilde{K}_0}{\epsilon} \mathrm{e}^{K_1 \epsilon^{\alpha/(\alpha+\beta)}} + 1 \leq \frac{K_0'}{\epsilon} \mathrm{e}^{K_1 \epsilon^{\alpha/(\alpha+\beta)}} ,$$

with total computational cost bounded as

$$\text{cost}(\hat{f}^{\text{CAIS}}_{h^*,N^*}) \leq \overline{\text{cost}}(\hat{f}^{\text{CAIS}}_{h^*,N^*}) \leq \frac{C^{\text{high}} K_0'}{\epsilon} \mathrm{e}^{K_1 \epsilon^{\alpha/(\alpha+\beta)}} + N_0 \epsilon^{-\beta/(\alpha+\beta)} .$$

□

Remark 6. Note that for when $c(h) = \beta^{1/h}$ and $\delta(h) = \alpha^{-1/h}$ in Theorem 2 we require that $h \in (0, \log(\alpha)/2]$ to ensure that $\delta(h)$ is convex. However, as $\epsilon \to 0$ we know that $\hat{h}^* \to 0$ as well, and therefore this requirement will automatically be satisfied for ϵ sufficiently small. The case where the tolerance ϵ is large so that the optimal fidelity $\hat{h}^* > \log(\alpha)/2$ corresponds to a very poor fidelity $\hat{h}^*$ being optimal.

The particular rates for the cost $c(h)$ and accuracy $\delta(h)$ of the surrogate densities $\pi^{(h)}$ commonly arise in Bayesian inverse problems (c.f. 1.3) and correspond to typical rates from numerical analysis where the fidelity h may correspond to a mesh width, for example. We now compare the upper bound on the cost complexity for the context-aware importance sampling estimator (2.39) derived in Theorem 2 with the cost complexity of the MFIS estimator (2.12) with a fixed fidelity $\bar{h} > 0$ that is independent of ϵ. For the fixed fidelity $\bar{h}$, the fewest number of samples $\bar{N}$ needed to guarantee that the MSE of the MFIS estimator $\hat{f}^{\mathrm{MFIS}}_{\bar{h},\bar{N}}$ is bounded above by ϵ is given by

$$\bar{N} = \left\lceil \frac{4 \left\| f \right\|^2_{L^\infty} \tilde{K}_0}{\epsilon} \mathrm{e}^{K_1 \delta(\bar{h})} \right\rceil .$$

The total costs (2.33) of the fixed fidelity estimator are therefore,

$$\mathrm{cost}(\hat{f}^{\mathrm{MFIS}}_{\bar{h},\bar{N}}) = \bar{N} C^{\mathrm{high}} + N_0 c(\bar{h}) . \tag{2.44}$$

In the first scenario of Theorem 2 where $\delta(h) = \alpha^{-1/h}$ and $\beta^{1/h}$, the costs (2.44) can be bounded above as

$$\mathrm{cost}(\hat{f}^{\mathrm{MFIS}}_{\bar{h},\bar{N}}) \leq \overline{\mathrm{cost}}(\hat{f}^{\mathrm{MFIS}}_{\bar{h},\bar{N}}) \leq \frac{C^{\mathrm{high}} K_0'}{\epsilon} \mathrm{e}^{K_1 \alpha^{-1/\bar{h}}} + N_0 \beta^{1/\bar{h}} .$$

Similarly, in the scenario where $\delta(h) = h^\alpha$ and $c(h) = h^{-\beta}$ the total costs are bounded as

$$\mathrm{cost}(\hat{f}^{\mathrm{MFIS}}_{\bar{h},\bar{N}}) \leq \overline{\mathrm{cost}}(\hat{f}^{\mathrm{MFIS}}_{\bar{h},\bar{N}}) \leq \frac{C^{\mathrm{high}} K_0'}{\epsilon} \mathrm{e}^{K_1 \bar{h}^\alpha} + N_0 \bar{h}^{-\beta} .$$

We now compare the cost complexity upper bound of the context-aware estimator $\overline{\mathrm{cost}}(\hat{f}^{\mathrm{CAIS}}_{h^*,N^*})$ to that of the standard MFIS estimator with a fixed fidelity $\overline{\mathrm{cost}}(\hat{f}^{\mathrm{MFIS}}_{\bar{h},\bar{N}})$ in the limit as the tolerance $\epsilon \to 0$. As $\epsilon \to 0$, the dominant term for the cost complexity of both estimators is due to the online phase whose costs scale as ϵ^{-1} and is the usual Monte Carlo cost-complexity

rate. For the fixed fidelity MFIS estimator the offline costs remain fixed with respect to ϵ and eventually become negligible relative to the online costs as $\epsilon \to 0$. On the other hand, for the CAIS estimator, the offline costs are adaptive to the tolerance ϵ since the online costs are also taken into consideration for the optimization problem (2.34). However, as $\epsilon \to 0$ the optimal fidelity $h^* \to 0$ as well, and hence the surrogate densities $\pi^{(h^*)}$ converge. This leads to diminishing returns for investing more computational resources to find better biasing densities at high fidelities (small h). Therefore, for the CAIS estimator the dominant term is also the online costs as $\epsilon \to 0$. When $\delta(h) = \alpha^{-1/h}$ and $c(h) = \beta^{1/h}$ we have that

$$\begin{aligned}
\lim_{\epsilon\to 0} \frac{\overline{\text{cost}}(\hat{f}^{\text{MFIS}}_{\bar{h},\bar{N}})}{\overline{\text{cost}}(\hat{f}^{\text{CAIS}}_{h^*,N^*})} &= \lim_{\epsilon\to 0} \frac{\frac{C^{\text{high}}K_0'}{\epsilon}\mathrm{e}^{K_1\alpha^{-1/\bar{h}}} + N_0\beta^{1/\bar{h}}}{\frac{C^{\text{high}}K_0'}{\epsilon}\mathrm{e}^{K_1\epsilon^{1/(1+\log_\alpha\beta)}} + N_0\epsilon^{-1/(1+\log_\beta\alpha)}} \\
&= \lim_{\epsilon\to 0} \frac{C^{\text{high}}K_0'\mathrm{e}^{K_1\alpha^{-1/\bar{h}}} + \epsilon N_0\beta^{1/\bar{h}}}{C^{\text{high}}K_0'\mathrm{e}^{K_1\epsilon^{1/(1+\log_\alpha\beta)}} + N_0\epsilon^{1-1/(1+\log_\beta\alpha)}} \\
&= \mathrm{e}^{K_1\alpha^{-1/\bar{h}}}.
\end{aligned}$$

Because $\mathrm{e}^{K_1\alpha^{-1/\bar{h}}} > 1$, as $\epsilon \to 0$ the CAIS estimator (2.39) obtains an asymptotic speedup over the MFIS estimator (2.12) with fixed fidelity $\bar{h}$. In particular, if $\bar{h}$ is small, corresponding to an accurate surrogate density $\pi^{(h)}$, the speedup factor becomes closer to 1 due to the diminishing returns of investing more resources into learning a better biasing density at higher fidelities. Similarly, we can derive an asymptotic speedup as $\epsilon \to 0$ for the CAIS estimator over the MFIS estimator at fidelity $\bar{h}$ when $\delta(h) = h^\alpha$ and $c(h) = h^{-\beta}$. As before we consider the ratio of the upper bounds on the cost complexity and take the limit $\epsilon \to 0$:

$$\begin{aligned}
\lim_{\epsilon\to 0} \frac{\overline{\text{cost}}(\hat{f}^{\text{MFIS}}_{\bar{h},\bar{N}})}{\overline{\text{cost}}(\hat{f}^{\text{CAIS}}_{h^*,N^*})} &= \lim_{\epsilon\to 0} \frac{\frac{C^{\text{high}}K_0'}{\epsilon}\mathrm{e}^{K_1\bar{h}^\alpha} + N_0\bar{h}^{-\beta}}{\frac{C^{\text{high}}K_0'}{\epsilon}\mathrm{e}^{K_1\epsilon^{\alpha/(\alpha+\beta)}} + N_0\epsilon^{-\beta/(\alpha+\beta)}} \\
&= \lim_{\epsilon\to 0} \frac{C^{\text{high}}K_0'\mathrm{e}^{K_1\bar{h}^\alpha} + \epsilon N_0\bar{h}^{-\beta}}{C^{\text{high}}K_0'\mathrm{e}^{K_1\epsilon^{\alpha/(\alpha+\beta)}} + N_0\epsilon^{1-\beta/(\alpha+\beta)}} \\
&= \mathrm{e}^{K_1\bar{h}^\alpha}.
\end{aligned}$$

Again we see that the speedup factor of the CAIS estimator over the MFIS estimator with fidelity $\bar{h}$ is strictly greater than 1 and moreover that smaller $\bar{h}$ again produces a smaller speedup. In both scenarios, $\delta(h) = \alpha^{-1/h}$ and $\delta(h) = h^{\alpha}$, the parameter α controls how quickly the surrogate densities $\pi^{(h)}$ converge to the high-fidelity target density π (c.f. Assumption 3) with larger α resulting in faster converging surrogate densities and a smaller asymptotic speedup. For general rates $\delta(h)$ the asymptotic speedup will depend on how quickly $\delta(h) \to 0$.

2.3.5 Computational procedure

Computing the context-aware estimator (2.39) involves the additional step of learning the context-aware surrogate model $\pi^{(h^*)}$, i.e. determining the optimal fidelity h^*, before the two-step process of computing the MFIS estimator (c.f. Section 2.2.3). Algorithm 2 summarizes the entire computational procedure for estimating a quantity of interest $\mathbb{E}_{\pi}[f]$ using context-aware importance sampling. Similar to other multifidelity methods, the procedure requires knowledge of the costs C^{high} and the function c for evaluating the unnormlaized high and low fidelity densities, respectively, as well the function δ from Assumption 3, which controls the accuracy of the surrogate densities. For context-aware importance sampling, the constants $\tilde{K}_0$ and K_1 from Theorem 2 must be provided as well. Note that for multifidelity importance sampling with a fixed fidelity $\bar{h}$, the constants $\tilde{K}_0$ and K_1 are still needed in order to provide guarantees on the mean-squared error, so the requirement is not unique to the CAIS estimator (2.39) only. Additionally, the tolerance ϵ and number of offline evaluations N_0 may be chosen by the user, although N_0 must be chosen to sufficiently fit the Laplace approximation (e.g. achieve machine precision in the gradient in the numerical optimization of (2.26)). The test function f must also be specified by the user as long as it is bounded with either the constant $\|f\|_{L^\infty}$ or some known upper bound on $|f|$. The algorithm first solves the relaxed optimization problem (2.35) to determine the fidelity h^* for the context-aware surrogate

model $\pi^{(h^*)}$ and number of online samples $N^* = \left\lceil \hat{N}^* \right\rceil$. Next the unnormalized surrogate density $\tilde{\pi}^{(h^*)}$ is used to fit the Laplace approximation μ_{h^*}. Since the Laplace approximation requires evaluating the Hessian at the mode we compute the mean of the Laplace approximation $\boldsymbol{\mu}_{h^*}^{\text{LAP}}$ using Newton's method, although any optimization method may be used to solve (2.26). Once the Hessian $\nabla\nabla^\top \Phi^{(h^*)} \boldsymbol{\mu}_{h^*}^{\text{LAP}}$ is computed we may either invert it to obtain the covariance of the Laplace approximation (2.27) or avoid the inversion and use it directly as the precision matrix, which is more amenable for high dimensional problems. Once the Laplace approximation μ_{h^*} has been computed, we draw $N^* = \left\lceil \hat{N}^* \right\rceil$ samples from μ_{h^*} and compute the importance weights (2.13) using the unnormalized high-fidelity target density $\tilde{\pi}$, to obtain the estimate (2.39).

Algorithm 2: Context-aware importance sampling

Input : Constants $\tilde{K}_0, K_1, C^{\text{high}}, \epsilon, N_0, \|f\|_{L^\infty}$ and functions c, δ

1 Solve the optimization problem (2.35) for $(\hat{h}^*, \hat{N}^*)$ using $\|f\|_{L^\infty}, \tilde{K}_0, K_1, C^{\text{high}}, N_0, \epsilon, c, \delta$;

2 Compute the Laplace approximation μ_{h^*} using $\tilde{\pi}^{(h^*)}$, where $h^* = \hat{h}^*$;

3 Draw $N^* = \left\lceil \hat{N}^* \right\rceil$ i.i.d. samples $\{\boldsymbol{\theta}^{[i]}\}_{i=1}^{N^*}$ from μ_{h^*};

4 Compute $\hat{f}_{h^*,N^*}^{\text{CAIS}}$ using (2.12);

Return: Estimate $\hat{f}_{h^*,N^*}^{\text{CAIS}}$

In practice, one may perform a pilot study to estimate the constants $C^{\text{high}}, \tilde{K}_0, K_1$ and functions c, δ that are not specified by the user and may be unknown ahead of time. Pilot studies are common in other multilevel and multifidelity methods where the algorithm parameters determine the trade-off between accuracy of the surrogate model and error of the estimator, as is the case here. Although the cost of the pilot study itself may be non-negligible it only needs to be performed once and may be amortized as the constants may be recycled to estimate arbitrary quantities of interest by replacing the test function f. Thus, context-aware importance sampling is attractive for estimating families of quantities of interest where the test function depends on an additional hyperparameter $f(\boldsymbol{\theta};\ \lambda)$ for $\lambda \in \Lambda$

such as cumulative density functions where

$$f(\theta;\ \lambda) = \mathbf{1}\{\boldsymbol{\theta} \leq \lambda\}, \quad \lambda \in \mathbb{R}\,,$$

or survival functions $1 - f(\theta; \lambda)$.

2.4 Bayesian inverse problems

Bayesian inverse problems, discussed in Section 1.3, are a prototypical example of where one can obtain surrogate densities $\pi^{(h)}$ by replacing the high-fidelity model G with a surrogate model $G^{(h)}$. In this section we show that the context-aware importance sampling estimator (2.39) can be applied to inference of a posterior arising from a Bayesian inverse problem where the target π is the posterior corresponding to the high-fidelity model. For Bayesian inference the posterior π is always absolutely continuous with respect to the prior π_0 and the χ^2 divergence $\chi^2\left(\pi \mid\mid \pi_0\right)$ of the posterior to the prior is bounded since

$$\int_{\Theta} \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} < \infty\,,$$

where we have used the fact that the likelihood $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$ is bounded with respect to $\boldsymbol{\theta}$. Therefore, one option to obtain a consistent estimator is to use importance sampling with the prior distribution as the biasing density. However, if the posterior concentrates around the observation $\mathbf{y}$, then $\chi^2\left(\pi \mid\mid \pi_0\right)$ may be large resulting in an estimator with high variance (c.f. Appendix B for linear inverse problems). A better option will be to use a biasing density that better adapts to the posterior such as the Laplace approximation or others discussed in Section 2.2.1.

Remark 7. Both a blessing and a curse: As discussed in Section 1.3.2, when the noise level goes to zero the posterior converges asymptotically to the Laplace approximation by

Bernstein-von Mises theorem [van der Vaart, 1998]. However because the posterior becomes more concentrated around its mean the optimization and computation of the Hessian becomes increasingly difficult minor numerical errors in either the mean or Hessian can result in poor sampling with large importance weights.

Recall from Section 1.3 that $G^{(h)}$ denotes the surrogate parameter-to-observable map with fidelity h and let it be such that the sequence $G^{(h)}(\boldsymbol{\theta}) \to G(\boldsymbol{\theta})$ converges pointwise for each $\boldsymbol{\theta} \in \Theta$. In addition to the set up in Section 1.3, we assume that $G, G^{(h)} \in \mathcal{C}^2(\Theta)$ to guarantee that the Hessian at the mode will exist for the Laplace approximation and moreover that the Taylor expansion in the proof of Proposition 1 will be valid. We consider the case where the prior π_0 is Gaussian $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, so that we can write the potential from Assumption 1 as

$$\Phi(\boldsymbol{\theta}) = \frac{1}{2} \left\| \mathbf{y} - G(\boldsymbol{\theta}) \right\|_{\Gamma^{-1}}^2 + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_0). \tag{2.45}$$

With a Gaussian prior, the resulting posterior distribution is always sub-Gaussian since we can take the matrix $\mathbf{A} = \frac{1}{4}\boldsymbol{\Sigma}_0^{-1}$ in Lemma 1. Without loss of generality assume that $\boldsymbol{\mu}_0 = \mathbf{0}$ so that

$$\mathrm{E}_{\pi_0} \left[\exp\left(\frac{1}{4} \boldsymbol{\theta}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\theta} \right) \right] \propto \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_0|^{1/2}} \int_{\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta};\ \mathbf{y}) \exp\left(-\frac{1}{4} \boldsymbol{\theta}^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\theta} \right) \mathrm{d}\boldsymbol{\theta} < \infty \,,$$

since the likelihood $\mathcal{L}(\boldsymbol{\theta};\ \mathbf{y})$ is bounded. The surrogate potentials $\Phi^{(h)}$ are defined similarly but with the surrogate maps $G^{(h)}$ replacing G.

2.4.1 Bounding chi-squared divergence with model error

In order to apply Theorem 1 and obtain a bound on the χ^2 divergence in terms of the model fidelity h, we translate bounds on the model error between G and $G^{(h)}$ to bounds on the

error between the corresponding potentials Φ and $\Phi^{(h)}$, respectively, for Assumption 3. The following two assumptions allow for this transition.

Assumption 5. The high-fidelity parameter-to-observable map G is globally Lipschitz meaning there exists a constant $B > 0$ such that for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$

$$\|G(\boldsymbol{\theta}) - G(\boldsymbol{\theta}')\| \leq B \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| .$$

Assumption 6. For all $\boldsymbol{\theta} \in \Theta$ and $h > 0$ we have

$$\left\|G^{(h)}(\boldsymbol{\theta}) - G(\boldsymbol{\theta})\right\| \leq \tilde{\delta}(h)\tilde{\tau}(\boldsymbol{\theta})$$

with $\tilde{\delta}(h) \to 0$ as $h \to 0$ with $\tilde{\tau}(\boldsymbol{\theta})$ independent of h.

Assumption 5 is almost the Lipschitz Assumption 2.7(ii) from [Stuart, 2010] except there the constant B only needs to hold for bounded sets of $\boldsymbol{\theta}$. Assumption 5 is satisfied if the map G is linear, for example, or if the map is the sum of a linear term and a smooth bounded function. Alternatively, we note that Assumption 5 may also be relaxed so that $G(\boldsymbol{\theta})$ grows at most linearly asymptotically as $\|\boldsymbol{\theta}\| \to \infty$. Such an assumption will still ensure that the potential does not grow faster than quadratically as needed for the assumptions of Theorem 1. Assumption 6 is similar to Assumption (4.11) in Corollary 4.9 of [Stuart, 2010], although the pointwise bound is also looser there than here because the χ^2 divergence is an upper bound on the Hellinger distance c.f. Remark 4. Theorem 3 is analogous to Theorem 1 from earlier but now is applied specifically to the Bayesian inverse problem.

Theorem 3. *If Assumptions 5 and 6 are satisfied with* $|\tilde{\tau}(\boldsymbol{\theta})| \leq \|\boldsymbol{\theta}\| + \tilde{\tau}_0$ *for some* $\tilde{\tau}_0 > 0$,

then Assumption 3 is also satisfied with

$$\delta(h) = \left(\frac{2B+1}{\kappa_{\min}}\right)\tilde{\delta}(h)$$

and $\tau(\boldsymbol{\theta})$ *a quadratic function of* $\|\boldsymbol{\theta}\|$ *that is independent of* h.

Proof. Using the form of the log-posterior (2.45) we write

$$\left|\Phi^{(h)}(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})\right| = \left|\|G^{(h)}(\boldsymbol{\theta}) - \mathbf{y}\|^2_{\Gamma^{-1}} - \|G(\boldsymbol{\theta}) - \mathbf{y}\|^2_{\Gamma^{-1}}\right|$$

since the prior terms cancel. To simplify notation, set $\Delta(\boldsymbol{\theta}) = G(\boldsymbol{\theta}) - G^{(h)}(\boldsymbol{\theta})$ and $\zeta(\boldsymbol{\theta}) = G(\boldsymbol{\theta}) - \mathbf{y}$, so that $\zeta(\boldsymbol{\theta}) - \Delta(\boldsymbol{\theta}) = G^{(h)}(\boldsymbol{\theta}) - \mathbf{y}$. Now, we can instead write

$$\begin{aligned}
\left|\Phi^{(h)}(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})\right| &= \left|\|\zeta(\boldsymbol{\theta})\|^2_{\Gamma^{-1}} - \|\zeta(\boldsymbol{\theta}) - \Delta(\boldsymbol{\theta})\|^2_{\Gamma^{-1}}\right| \\
&= \left|\|\zeta(\boldsymbol{\theta})\|^2_{\Gamma^{-1}} - \left\langle \Gamma^{-1}\left(\zeta(\boldsymbol{\theta}) - \Delta(\boldsymbol{\theta})\right),\ \zeta(\boldsymbol{\theta}) - \Delta(\boldsymbol{\theta})\right\rangle\right| \\
&= \left|2\langle \Delta(\boldsymbol{\theta}), \Gamma^{-1}\zeta(\boldsymbol{\theta})\rangle - \|\Delta(\boldsymbol{\theta})\|^2_{\Gamma^{-1}}\right| .
\end{aligned}$$

Applying the triangle inequality and then the Cauchy-Schwarz inequality to this last line gives

$$\left|\Phi^{(h)}(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})\right| \leq 2\|\Delta(\boldsymbol{\theta})\|\|\Gamma^{-1}\zeta(\boldsymbol{\theta})\| + \|\Delta(\boldsymbol{\theta})\|^2_{\Gamma^{-1}} . \tag{2.46}$$

Using that $\mathbf{y} = G(\boldsymbol{\theta}^\star) + \boldsymbol{\eta}$ and the triangle inequality gives

$$\begin{aligned}
\|\Gamma^{-1}\zeta(\boldsymbol{\theta})\| &= \|\Gamma^{-1}(G(\boldsymbol{\theta}) - \mathbf{y})\| \\
&\leq \|\Gamma^{-1}(G(\boldsymbol{\theta}) - G(\boldsymbol{\theta}^\star))\| + \|\Gamma^{-1}\boldsymbol{\eta}\| .
\end{aligned}$$

Assumption 5 then gives the bound

$$\|\Gamma^{-1}\zeta(\boldsymbol{\theta})\| \leq \frac{B}{\kappa_{\min}}\|\boldsymbol{\theta}-\boldsymbol{\theta}^{\star}\| + \|\Gamma^{-1}\boldsymbol{\eta}\|, \tag{2.47}$$

where $\kappa_{\min} > 0$ is the smallest eigenvalue of the covariance matrix Γ, i.e., the direction along which the posterior is most peaked. Similarly, we bound

$$\|\Delta(\boldsymbol{\theta})\|^2_{\Gamma^{-1}} = \langle \Gamma^{-1}\Delta(\boldsymbol{\theta}),\ \Delta(\boldsymbol{\theta})\rangle \leq \frac{1}{\kappa_{\min}}\|\Delta(\boldsymbol{\theta})\|^2\,. \tag{2.48}$$

Substituting bounds (2.47) and (2.48) into (2.46) yields

$$\left|\Phi^{(h)}(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})\right| \leq 2\left(\frac{B}{\kappa_{\min}}\left(\|\boldsymbol{\theta}-\boldsymbol{\theta}^{\star}\|\right) + \|\Gamma^{-1}\boldsymbol{\eta}\|\right)\|\Delta(\boldsymbol{\theta})\| + \frac{1}{\kappa_{\min}}\|\Delta(\boldsymbol{\theta})\|^2,$$

and the triangle inequality gives

$$\left|\Phi^{(h)}(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})\right| \leq 2\left(\frac{B}{\kappa_{\min}}\left(\|\boldsymbol{\theta}\| + \|\boldsymbol{\theta}^{\star}\|\right) + \|\Gamma^{-1}\boldsymbol{\eta}\|\right)\|\Delta(\boldsymbol{\theta})\| + \frac{1}{\kappa_{\min}}\|\Delta(\boldsymbol{\theta})\|^2\,. \tag{2.49}$$

Assumption 6 along with the assumption that $|\tilde{\tau}(\boldsymbol{\theta})| \leq \|\boldsymbol{\theta}\| + \tilde{\tau}_0$ says $\|\Delta(\boldsymbol{\theta})\| \leq \tilde{\delta}(h)\left(\|\boldsymbol{\theta}\| + \tilde{\tau}_0\right)$, so we get that

$$\left|\Phi^{(h)}(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})\right| \leq 2\left(\frac{B}{\kappa_{\min}}\left(\|\boldsymbol{\theta}\| + \|\boldsymbol{\theta}^{\star}\|\right) + \|\Gamma^{-1}\boldsymbol{\eta}\|\right)\tilde{\delta}(h)\left(\|\boldsymbol{\theta}\| + \tilde{\tau}_0\right) + \frac{1}{\kappa_{\min}}\tilde{\delta}(h)^2\left(\|\boldsymbol{\theta}\| + \tilde{\tau}_0\right)^2\,,$$

and thus

$$\begin{aligned}\left|\Phi^{(h)}(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})\right| \leq &\left(\frac{2B}{\kappa_{\min}}\|\boldsymbol{\theta}^\star\| + 2\|\Gamma^{-1}\boldsymbol{\eta}\| + \frac{\tilde{\delta}(h)\tilde{\tau}_0}{\kappa_{\min}}\right)\tilde{\delta}(h)\tilde{\tau}_0 \\ &+ \left(\frac{2B}{\kappa_{\min}}\tilde{\tau}_0 + \frac{2B}{\kappa_{\min}}\|\boldsymbol{\theta}^\star\| + \frac{2}{\kappa_{\min}}\tilde{\delta}(h)\tilde{\tau}_0 + 2\|\Gamma^{-1}\boldsymbol{\eta}\|\right)\tilde{\delta}(h)\|\boldsymbol{\theta}\| \\ &+ \left(\frac{2B}{\kappa_{\min}} + \frac{1}{\kappa_{\min}}\tilde{\delta}(h)\right)\tilde{\delta}(h)\|\boldsymbol{\theta}\|^2 .\end{aligned}$$

Using that $\tilde{\delta}(h) \leq 1$ for all h sufficiently small and $\|\boldsymbol{\theta}\| \leq 1 + \|\boldsymbol{\theta}\|^2$ gives

$$\left|\Phi^{(h)}(\boldsymbol{\theta}) - \Phi(\boldsymbol{\theta})\right| \leq \delta(h)\tau(\boldsymbol{\theta}),$$

where

$$\delta(h) = \left(\frac{2B+1}{\kappa_{\min}}\right)\tilde{\delta}(h)$$

is as in Assumption 3 and $\tau(\boldsymbol{\theta})$ is quadratic in $\|\boldsymbol{\theta}\|$ and is bounded independent of h. □

Corollary 1. *Suppose that Theorem 1 applies with Assumption 3 provided by Theorem 3. Then, together with Proposition 1 this implies that the cost complexity of the context-aware importance sampling estimator with a Laplace approximation biasing density is given by Theorem 2.*

2.5 Numerical results

The following three numerical examples demonstrate the context-aware importance sampling estimator (2.39) and in particular its cost complexity compared with the standard MFIS estimator (2.12) at a fixed fidelity $\bar{h}$. All runtime measurements were performed on compute nodes equipped with Intel Xeon Gold 6148 2.4GHz processors and 192GB of memory using a Python 3.6 implementation.

2.5.1 Steady-state heat conduction

In this section we consider the problem of inferring the heat-conductivity of a one-dimensional rod whose temperature is governed by a steady-state heat equation.

Problem Setup

Let $\Omega = (0,1) \subset \mathbb{R}$ and $\Theta = \mathbb{R}^6$ and consider the parametric differential equation for the temperature $u : \Omega \times \Theta \to \mathbb{R}$

$$\begin{aligned} -\left(\exp\left(\kappa(x;\boldsymbol{\theta})\right) u_x(x;\boldsymbol{\theta})\right)_x &= 1\,, \quad x \in \Omega \\ u(0;\boldsymbol{\theta}) &= 0\,, \\ k(1;\boldsymbol{\theta}) u_x(1;\boldsymbol{\theta}) &= 0\,, \end{aligned} \tag{2.50}$$

parameterized by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_6)^\top \in \Theta$ and where $\kappa : \Omega \times \Theta \to \mathbb{R}$ is the log heat conductivity. The log heat conductivity $\kappa(x;\ \boldsymbol{\theta})$ is a smoothed piecewise constant function over the interval $[0,1]$ such that $\kappa(x;\boldsymbol{\theta}) \approx \theta_i$ over the sub-interval $x \in \left[\frac{i-1}{6},\ \frac{i}{6}\right)$ for $i = 1, \ldots, 6$. In particular, let

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

and $\alpha_i = (i-1)/6$ for $i = 1, \ldots, 6$. Define

$$\hat{\kappa}_i(x;\boldsymbol{\theta}) = (1 - I(x,\alpha_i))\hat{\kappa}_{i-1}(x;\boldsymbol{\theta}) + I(x,\alpha_i)\theta_i \tag{2.51}$$

for $i = 2, \ldots, 6$ and $\hat{\kappa}_1(x;\boldsymbol{\theta}) = \theta_1$, and set $\kappa = \hat{\kappa}_6$. Figure 2.2 illustrates a sample log heat-conductivity with the quantity 0.005 in (2.51) controlling the width of the smoothing between the sub-intervals. For brevity, define $\mathcal{I}^{\text{int}} : \mathbb{R}^6 \to \mathcal{C}[0,1]$ to be the interpolation operator that maps the parameter $\boldsymbol{\theta} \in \mathbb{R}^6$ to the continuous function $\kappa(\cdot;\boldsymbol{\theta}) : [0,1] \to \mathbb{R}$ given by (2.51). The high-fidelity and surrogate models G and $G^{(h)}$ discretize (2.50) in the

spatial domain Ω using linear finite elements with an equi-spaced mesh of width $h > 0$ to obtain an approximation $u^{(h)}$. The discretization gives rise to a sparse tri-diagonal linear system which is positive definite due to the positivity of the heat conductivity $\exp(\kappa(x;\boldsymbol{\theta}))$ and is solved using SciPy's solveh_banded method that wraps LAPACK's pbsv function, which performs a Cholesky factorization of the system. Let $F^{(h)} : C^1[0,1] \to C[0,1]$ denote the solution operator that maps $\kappa(\cdot;\boldsymbol{\theta})$ to the finite element solution $u^{(h)}(\cdot,\boldsymbol{\theta})$. The finite element solution $u^{(h)}$ is then observed at 120 equally-spaced points throughout Ω with the observation operator $\mathcal{B}^{\text{obs}} : C[0,1] \to \mathbb{R}^{120}$ defined by

$$\mathcal{B}^{\text{obs}}(u^{(h)})_i = u^{(h)}(x_i;\boldsymbol{\theta}), \quad x_i = \frac{i}{120},$$

for $i = 1, \ldots, 120$. The full parameter-to-observable surrogate models are then given by

$$G^{(h)} = \mathcal{B}^{\text{obs}} \circ F^{(h)} \circ \mathcal{I}^{\text{int}},$$

for $h_i^{-1} = 4(i+1)$ with $i = 1, \ldots, 15$ (multiples of 4 for the number of elements in the mesh) and the high-fidelity model is given by using $h_0^{-1} = 256$ elements

$$G = \mathcal{B}^{\text{obs}} \circ F^{(h_0)} \circ \mathcal{I}^{\text{int}}.$$

Setup of the inverse problem

Let $\mathbf{y} \in \mathbb{R}^{120}$ be a single observation given by evaluating the high-fidelity model G at the true parameter $\boldsymbol{\theta}^\star = \mathbf{1}_6$ (a 6-dimensional vector of all ones) and perturbed by Gaussian noise

$$\mathbf{y} = G(\boldsymbol{\theta}^*) + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim N(\mathbf{0}, 10^{-5}\mathbf{I}_{120\times 120}).$$

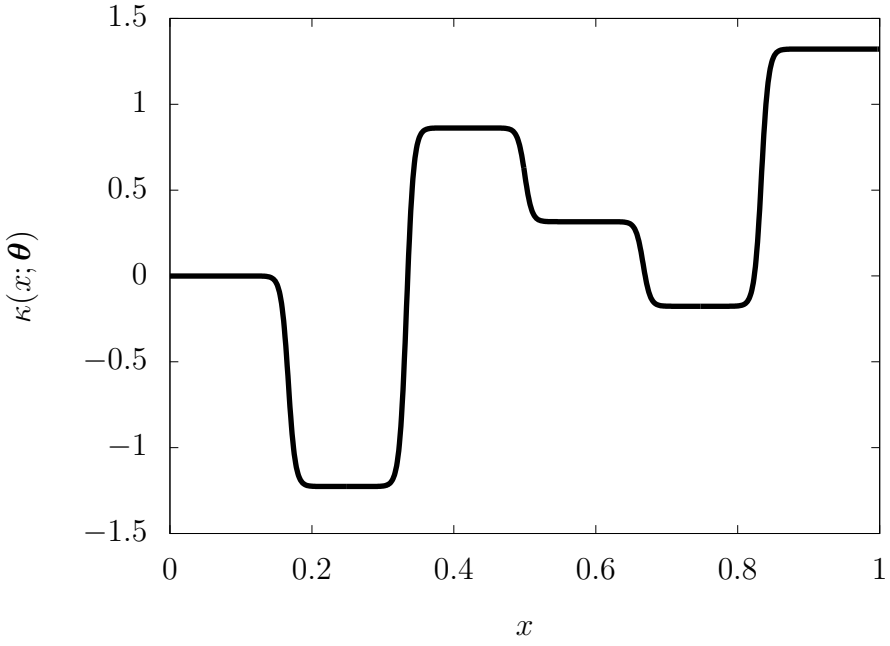


Figure 2.2: A smoothed piecewise constant function given by (2.51).

The standard deviation $\sqrt{10^{-5}}$ of the added noise $\boldsymbol{\eta}$ corresponds to approximately 1% of the high-fidelity solution $u = F^{(h_0)} \circ \mathcal{I}^{\text{int}}(\boldsymbol{\theta}^\star)$ at the right endpoint $x = 1$. The true solution $u(x;\boldsymbol{\theta}^\star)$ and the observed data $\mathbf{y}$ are show in Figure 2.3.

The prior distribution is taken to be Gaussian with mean $\boldsymbol{\mu}_0 = \mathbf{1}_6$ and prior covariance $\boldsymbol{\Sigma}_0 = 10^{-1}\mathbf{I}_{6\times 6}$. The test function f for the quantity of interest (2.1) is taken to be

$$f(\boldsymbol{\theta}) = 2 \cdot \mathbf{1}\{(\boldsymbol{\theta} - \boldsymbol{\mu}^{\text{LAP}})^\top \mathbf{v}_1 \geq 0\} - 1\,, \tag{2.52}$$

where $\boldsymbol{\mu}^{\text{LAP}}$ and $\boldsymbol{\Sigma}^{\text{LAP}}$ are the mean and covariance of the Laplace approximation to the high-fidelity posterior $\pi = \pi^{(h_0)}$ and $\mathbf{v}_1$ is the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\Sigma}^{\text{LAP}}$ normalized to have unit norm. The choice (2.52) for the test function f is motivated by trying to maximize the left-hand-side of the inequality

$$(f - \mathbb{E}_\pi[f])^2 \leq 4\,\|f\|_{L^\infty}\,,$$

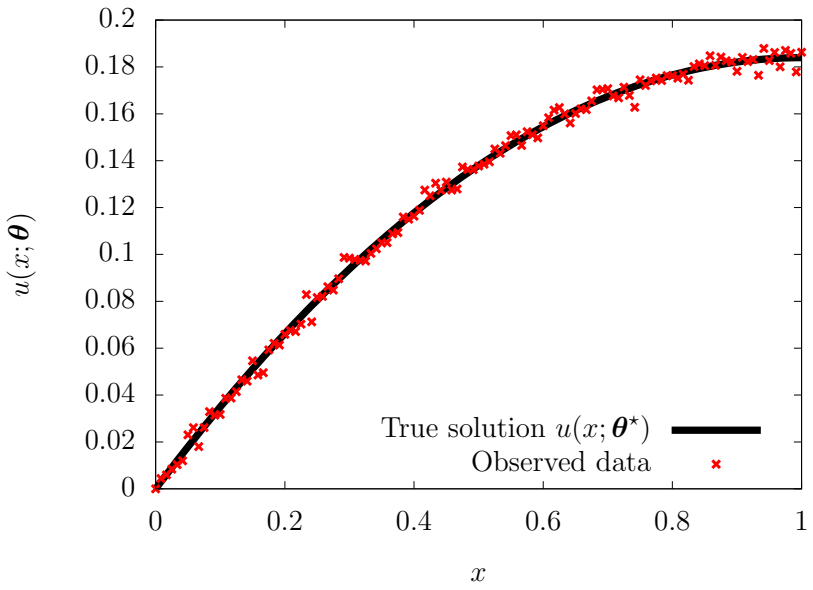


Figure 2.3: The true solution $u(x;\boldsymbol{\theta}^\star)$ of the steady-state heat equation (2.50) and the observed data $\mathbf{y} \in \mathbb{R}^{120}$ over the domain.

so that the bound (2.8) on the mean-squared error of the MFIS estimator (2.12) is as tight as possible. Since the test function (2.52) satisfies $f(\boldsymbol{\theta}) \in \{\pm 1\}$ for all $\boldsymbol{\theta} \in \Theta$, in the case where the high-fidelity posterior π is exactly Gaussian so that the Laplace approximation $\pi = \mu_{h_0}$ we know that by symmetry $\mathbb{E}_\pi[f] = 0$ and therefore

$$(f - \mathbb{E}_\pi[f])^2 = \|f\|_{L^\infty} \ .$$

Pilot study

To estimate the constants $\tilde{K}_0$ and K_1 required for Algorithm 2 in computing the CAIS estimator (2.39) we perform a pilot study in which we first estimate the χ^2 divergences from the Laplace approximations μ_{h_i} of the surrogate densities $\pi^{(h_i)}$ for $i = 1, \ldots, 15$ to the high-fidelity target density π and then fit a curve for the upper bound (2.24). First we note that because the solution operator $F^{(h)}$ approximates the solution u to (2.50) with linear

finite elements the accuracy of the surrogate models has the form $\delta(h) = h^2$ for δ as in Assumption 3. Moreover, the costs may be modeled as

$$c(h) = c_0 + \frac{c_1}{h}\,,$$

which scales linearly with the number of elements h^{-1} due to the sparse tri-diagonal structure of the resulting linear system. The Laplace approximation μ_{h_i} to each surrogate posterior $\pi^{(h_i)}$ for $i = 1, \ldots, 15$ is fit using Newton's method until the norm of the gradient $\left\|\nabla\Phi^{(h_i)}\right\|$ has reached machine precision. In this problem the gradient and Hessian matrix are computed using a second-order finite difference scheme and find that 15 Newton iterations with $N_0 = 1150$ total model evaluations are sufficient to fit the Laplace approximation for each fidelity. Once the Laplace approximations μ_{h_i} for $i = 1, \ldots, 15$ have been fit we estimate the χ^2 divergence with Monte Carlo estimator

$$\hat{\chi}^2_{h,N} = N\frac{\sum_{i=1}^N \left(\tilde{\pi}(\boldsymbol{\theta}^{[i]})/\mu_h(\boldsymbol{\theta}^{[i]})\right)^2}{\left(\sum_{i=1}^N \tilde{\pi}(\boldsymbol{\theta}^{[i]})/\mu_h(\boldsymbol{\theta}^{[i]})\right)^2}\,, \quad \{\boldsymbol{\theta}^{[i]}\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \mu_h\,. \tag{2.53}$$

As $N \to \infty$ the estimator (2.53) converges almost surely to $\chi^2\left(\pi \,||\, \mu_h\right) + 1$. Here we compute $\hat{\chi}^2_{h,N}$ with $N = 10^3$ samples for each fidelity h_i, $i = 1, \ldots, 15$ and average over $N_{\text{rep}} = 500$ independent trials so that

$$\hat{\chi}^2_{\text{meas},h} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left(\hat{\chi}^2_{h,N}\right)^{[i]}\,, \tag{2.54}$$

where the index i corresponds to an independent trial of estimating (2.53). Thus, a total of 5×10^5 samples are drawn. The constants $\tilde{K}_0$ and K_1 are then estimated by fitting a curve of the form

$$\tilde{K}_0 \mathrm{e}^{K_1\delta(h)}\,,$$

where recall that $\delta(h) = h^2$, to the data points $\{(h_i, \hat{\chi}^2_{\text{meas},h_i})\}_{i=1}^{15}$ using a log transformation and then fitting a first-order polynomial to

$$\log(\tilde{K}_0) + K_1\delta(h)\,.$$

Results

Figure 2.4 shows the fitted curve using the constants $\tilde{K}_0$ and K_1 obtained during the pilot study along with the estimated χ^2 data $\{(h_i, \hat{\chi}^2_{\text{meas},h_i})\}_{i=1}^{15}$. For low fidelities, i.e. large mesh width h, the χ^2 divergence is large due to the poor approximation of the surrogate density. As the fidelity improves and $h \to 0$, i.e. number of elements increases, the χ^2 divergence quickly converges to a constant that depends on the use of the Laplace approximation as the biasing density rather than the surrogate density itself. If a more flexible family of biasing densities were considered such as Gaussian mixtures, transport maps, or normalizing flows, then the limiting constant may be brought down even further. Since we only consider finitely many surrogate models $_1, \ldots, h_15$, we approximate the solution of the optimization problem (2.34) with a brute force search to find the best fidelity $h^* \in \{h_1, \ldots, h_{15}\}$ from the list of fidelities that we consider and set $N^* = \left\lceil \hat{N}^* \right\rceil$ with $\hat{N}^*$ corresponding to h^* through

$$\hat{N}^* = \frac{4\,\|f\|^2_{L^\infty}\,\tilde{K}_0}{\epsilon}\mathrm{e}^{K_1\delta(h^*)}\,. \tag{2.55}$$

Figure 2.4 shows the selected fidelity h^* and thus the corresponding optimal number of elements as a function of the tolerance ϵ on the MSE. As the tolerance shrinks a higher fidelity model is required to fit the Laplace approximation to reduce the number of samples needed in the online phase. Furthermore, notice that because the χ^2 divergence $\chi^2\left(\pi \,||\, \mu_h 1\right)$ where $h_1^{-1} = 8$ is significantly larger than the rest that this surrogate model is never selected, even when the tolerance is very large. Figure 2.5 shows the theoretical optimal trade-off between

cost (runtime in seconds) and the MSE (tolerance) of the CAIS estimator (2.39) compared with the trade-off for MFIS estimator (2.12) with a fixed fidelity. Note that the costs of the pilot study are not included in the total costs of computing the CAIS estimator (2.39) as the constants are assumed to be provided as in Algorithm 2. Moreover, we assume that the cost of solving the optimization problem (2.35) for the optimal fidelity is itself negligible relative to the cost of evaluating the surrogate and high-fidelity models. For the actual MSE of the estimators, first a ground truth reference value $\bar{f}$ for $\mathbb{E}_\pi[f]$ was computed using 10^5 evaluations of the high-fidelity model with the estimator $\hat{f}^{\text{MFIS}}_{h_0,10^5}$ and averaged the results over $N_2 = 500$ independent trials (again denoted by the superscript $[i]$) for 5×10^7 total samples

$$\bar{f} = \frac{1}{N_2} \sum_{i=1}^{N_2} \hat{f}^{[i]}_{h_0,10^5} \,. \tag{2.56}$$

Then, for each tolerance ϵ the MSE of the CAIS estimator (2.39) is estimated using $N_3 = 1000$ trials

$$\widehat{\text{MSE}}_\epsilon = \frac{1}{N_3} \sum_{i=1}^{N_3} \left(\left(\hat{f}^{\text{CAIS}}_{h^*,N^*} \right)^{[i]} - \bar{f} \right)^2 , \tag{2.57}$$

where the subscript ϵ denotes the dependence of the pair (h^*, N^*), and thus the CAIS estimator, on the tolerance ϵ. Figure 2.5 shows both the estimated MSE (2.57) over $N_3 = 1000$ trials for different tolerances ϵ as well as the MSE for the estimators $\hat{f}_{h_0,N(h_0)}$ and $\hat{f}_{h_1,N(h_1)}$ where the number of samples depending on h is

$$N(h) = \left\lceil \frac{\tilde{K}_0}{\epsilon} \exp\left(K_1 h^2\right) \right\rceil$$

and $h_1 = 8$ is the lowest fidelity we consider (for the surrogate only estimator we average only $N_3 = 500$ trials). Figure 2.5 shows the estimated MSE (2.57) over $N_3 = 1000$ trials for the CAIS estimator (2.39) as well as two MFIS estimators with fixed fidelities $h_0 = 1/256$ and $h_1 = 1/8$, respectively the highest and lowest fidelities. Note that the true MSE estimated

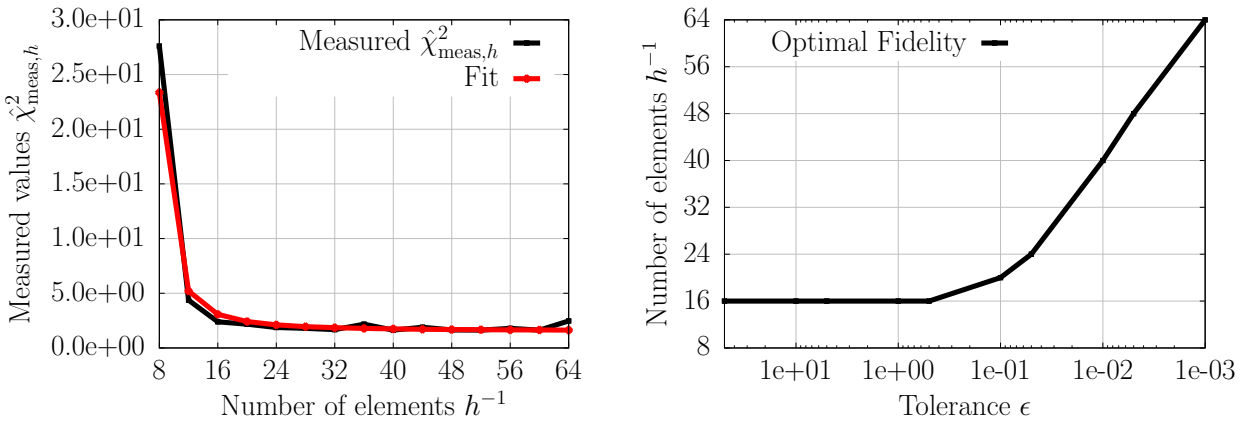


Figure 2.4: (**Left**) The measured χ^2 divergences $\hat{\chi}^2_{\text{meas},h}$ given by (2.54) of $\chi^2(\pi \,||\, \mu_{h_i})$ for $i = 1, \ldots, 15$ and the corresponding fitted curve. (**Right**) The selected fidelity $h^* \in \{h_1, \ldots, h_{15}\}$ for the optimal number of elements $(h^*)^{-1}$ from the optimization (2.34) as the tolerance ϵ on the MSE changes.

in the right plot of 2.5 is bounded above by the tolerance ϵ shown in the left plot due to the constraint in the optimization problem (2.34). For moderate error tolerances as small as $\epsilon = 10^{-2}$ we observe an order of magnitude speedup of the CAIS estimator (2.39) over the MFIS estimator with the fidelity fixed to be the high-fidelity h_0. This speedup is attributed to the reduction in cost of fitting the Laplace approximation to the cheaper surrogate density $\pi^{(h^*)}$. For these moderate error tolerances, few samples are needed in the online phase to sufficiently reduce the variance and thus using the high-fidelity density π to learn a more accurate biasing density is excessive. Conversely, as the tolerance $\epsilon \to 0$, most of the computational costs shift to the online sampling phase punishing poor biasing densities derived from low fidelity surrogate models (e.g. $\pi^{(h_1)}$ in Figure 2.5) and resulting in little speedup. Since here we consider only a fixed set of fidelities $h_0 > h_1 > \ldots > h_{15}$, once the tolerance ϵ is sufficiently small the high-fidelity density $\pi = \pi^{(h_0)}$ will be selected and no more speedup will be achieved as one may consider $\delta(h) = 0$ for $h \leq h_0$. These empirical observations are consistent with the theoretical speedup derived in Section 2.3.4, which in particular depended upon the rate at which $\delta(h) \to 0$.

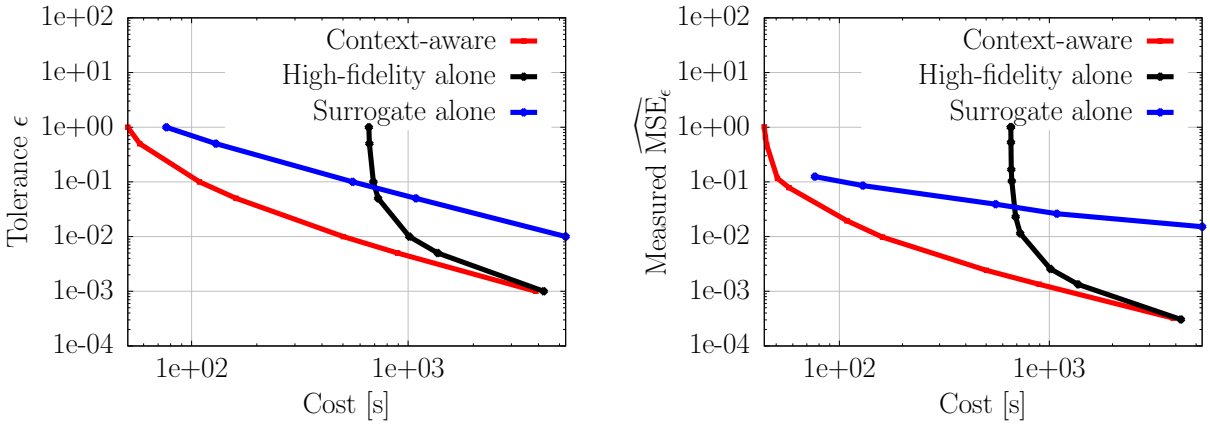


Figure 2.5: (**Left**) The error tolerance ϵ against the total costs (2.33) (runtime in seconds) to fit the Laplace approximation μ_{h^*} of $\pi^{(h^*)}$ and compute the CAIS estimator $\hat{f}^{\text{CAIS}}_{h^*,N^*}$ with N^* samples compared against the MFIS estimator (2.12) with a fixed fidelity. (**Right**) The estimated MSE, $\widehat{\text{MSE}}_\epsilon$, with respect to ground truth reference value against the total costs.

2.5.2 Euler Bernoulli Beam Model

In this example, we infer the effective stiffness of an Euler Bernoulli beam and follow the setup presented in Section 4.2 of [Peherstorfer and Marzouk, 2019]. The forward-model code in this example is available on GitHub[1] and was developed by Matthew Parno as a part of the 2018 Gene Golub SIAM Summer School on "Inverse Problems: Systematic Integration of Data with Models under Uncertainty".

Problem Setup

Consider a cantilever beam of unit length modeled by the unit interval $\Omega = (0, 1) \subset \mathbb{R}$ and fixed at the origin. Given an applied force $g(x)$ throughout the domain Ω the displacement $u : \Omega \times \Theta \to \mathbb{R}$ of the beam is governed by the parametric fourth-order differential equation

$$\frac{\partial^2}{\partial x^2}\left(E(x;\boldsymbol{\theta})\frac{\partial^2}{\partial x^2}u(x;\boldsymbol{\theta})\right) = g(x), \quad x \in \Omega \tag{2.58}$$

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

with boundary conditions

$$u(0;\boldsymbol{\theta}) = 0, \quad \frac{\partial u}{\partial x}(0;\boldsymbol{\theta}) = 0, \quad \frac{\partial^2 u}{\partial x^2}(1;\boldsymbol{\theta}) = 0, \quad \frac{\partial^3 u}{\partial x^3}(1;\boldsymbol{\theta}) = 0$$

where $\Theta = \mathbb{R}^6$ and $E : \Omega \times \Theta \to \mathbb{R}$ is the effective stiffness of the beam. Here $g(x) = 1$ is taken to be constant and so the effective stiffness function $E(x;\boldsymbol{\theta})$ completely determines the displacement of the beam with larger effective stiffness resulting in less displacement. As in the steady-state heat condutction example 2.5.1, the effective stiffness is taken to be a smoothed piecewise constant function given by (2.51) but with the parameter coordinates $\boldsymbol{\theta}_i$ replaced by their absolute values $|\theta_i|$ for $i = 1, \ldots, 6$ to enforce positivity of the effective stiffness. Again we let $\mathcal{I}^{\text{int}} : \mathbb{R}^6 \to C^2[0,1]$ denote the interpolation operator that maps the parameters $\boldsymbol{\theta}$ to the effective stiffness function $E(\cdot;\boldsymbol{\theta})$. Note that taking the absolute values $|\theta_i|$ of the parameters in (2.51) will result in a parameter-to-observable map that is not differentiable. However, in the set up of the inverse problem, the prior, and therefore posterior, will be concentrated in a region of Θ away from the set

$$\{\boldsymbol{\theta} \in \Theta : \exists i \in \{1, \ldots, 6\}, \theta_i = 0\},$$

where the interpolation operator is not differentiable. In other words, the parameter-to-observable map will be smooth in the region of the prior and posterior and hence fitting the Laplace approximation is not an issue. Given an effective stiffness $E(x;\boldsymbol{\theta})$, the forward models $F^{(h)}$ solve the equation (2.58) using a second-order finite difference scheme with a mesh width $h > 0$ that gives $h^{-1} + 1$ total grid points $x_i = ih$ for $i = 0, \ldots, h^{-1}$. The resulting sparse linear system of equations is then solved using SciPy's spsolve function to obtain a vector of the approximate solution $\mathbf{u}^{(h)} \in \mathbb{R}^{h^{-1}+1}$ evaluated at the grid points $x_0, \ldots, x_{h^{-1}}$. Finally, the solution operator $F^{(h)} : C^2[0,1] \to C[0,1]$ maps the finite difference solution vector $\mathbf{u}^{(h)}$ to its continuous piecewise linear interpolant $u^{(h)} : [0,1] \to \mathbb{R}$ such that

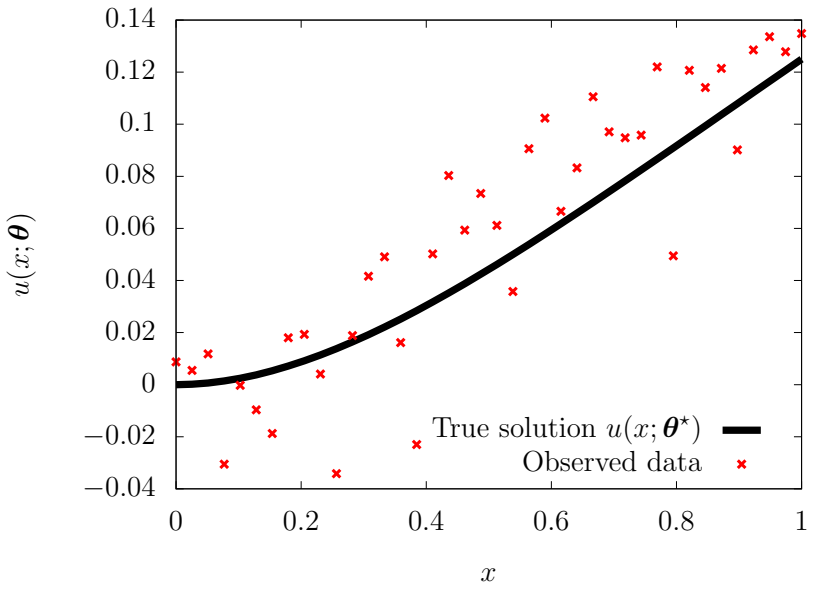


Figure 2.6: The true solution $u(x;\boldsymbol{\theta}^\star)$ of the Euler-Bernoulli equation (2.58) and the observed data $\mathbf{y} \in \mathbb{R}^{40}$ over the domain.

$u^{(h)}(x_i) = (\mathbf{u}^{(h)})_i$ for $i = 0, \ldots, h^{-1}$. The approximate solution $u^{(h)}$ is then observed at 40 equally spaced points throughout the unit interval with the observation operator $\mathcal{B}^{\text{obs}} : C[0,1] \to \mathbb{R}^{40}$ defined by

$$\mathcal{B}^{\text{obs}}(u^{(h)})_i = u^{(h)}\left(\frac{i-1}{39}\right), \quad i = 1, \ldots, 40\,,$$

so that the full parameter-to-observable maps are given by $G^{(h)} = \mathcal{B}^{\text{obs}} \circ F^{(h)} \circ \mathcal{I}^{\text{int}} : \mathbb{R}^6 \to \mathbb{R}^{40}$. Note that the left end-point at $x = 0$ is not observed since it is fixed by the boundary condition $u(0;\boldsymbol{\theta}) = 0$. As with the steady-state heat conduction problem, the high-fidelity model G is chosen to use $h_0^{-1} + 1 = 256$ grid points and the surrogate models $G^{(h)}$ use $h_i^{-1} + 1 = 4(i+1)$ grid points for $i = 1, \ldots, 15$.

Setup of the inverse problem

A single synthetic observation $\mathbf{y} \in \mathbb{R}^{40}$ by evaluating the ground truth parameters $\boldsymbol{\theta}^\star = \mathbf{1}_6 \in \mathbb{R}^6$ and perturbing with Gaussian noise

$$\mathbf{y} = G(\boldsymbol{\theta}^\star) + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{\Gamma}),$$

with noise covariance $\boldsymbol{\Gamma} = 5.623 \times 10^{-4} \mathbf{I}_{40\times 40}$. The standard deviation of the noise $\sqrt{5.623 \times 10^{-4}}$ corresponds to 5% of the true solution $u = F^{(h_0)} \circ \mathcal{I}^{\text{int}}(\boldsymbol{\theta}^\star)$ at the end of the beam $x = 1$. The prior π_0 is taken to be Gaussian with mean $\boldsymbol{\mu}_0 = \mathbf{1}_6$ and covariance $\boldsymbol{\Sigma}_0 = 1.778 \times 10^{-2} \mathbf{I}_{6\times 6}$. The test function f for the quantity of interest (2.1) is the same as in (2.52) where we recall that its purpose was to obtain as tight of an upper bound for (2.8) as possible.

Pilot study

We perform a similar pilot study as in 2.5.1 to estimate the constants $\tilde{K}_0, K_1$ required for computing Algorithm 2. Again we fit a Laplace approximation μ_{h_i} for each surrogate density $\pi^{(h_i)}$, $i = 1, \ldots, 15$ using Newton's method with finite difference approximations for the gradient and Hessian at each iteration. We find that 20 iterations of Newton's method with the finite difference approximations for the derivatives is sufficient to achieve near machine precision in the norm of the gradient $\nabla \Phi^{(h_i)}$ for each surrogate model. Thus, when combined with the additional model evaluations to approximate the Hessian, a total of $N_0 = 1800$ surrogate model evaluations are performed throughout the entire offline phase. Because a second-order finite difference scheme is used to discretize (2.58), the accuracy of the surrogate models have the form $\delta(h) = h^2$. Moreover, the stiffness matrix for the resulting linear system is sparse so the costs $c(h)$ scale linearly in the number of grid points and we model

$$c(h) = c_0 + \frac{c_1}{h}.$$

The χ^2 divergences $\chi^2(\pi \,||\, \mu_{h_i})$ for $i = 1, \ldots, 15$ are then estimated with estimator $\hat{\chi}^2_{h_i,10^5}$ from (2.53) and averaged over $N_1 = 100$ independent trials to obtain the estimated value $\hat{\chi}^2_{\text{meas},h_i}$ as in (2.54). The constants $\tilde{K}_0$ and K_1 are then fit to the data $\{(h_i, \hat{\chi}^2_{\text{meas},h_i})\}_{i=1}^{15}$ using a linear polynomial for

$$\log(\tilde{K}_0) + K_1 h^2 \,.$$

Results

Figure 2.7 shows both the estimated χ^2 values $\hat{\chi}^2_{\text{meas},h_i}$ as well as the fitted curve $\tilde{K}_0 \mathrm{e}^{K_1 h^2}$. As in the steady-state heat conduction problem 2.5.1, the χ^2 divergence quickly levels off once the number of grid points is sufficient to provide a indicating an accurate biasing density. The lowest fidelity surrogate models are never selected due to the large χ^2 divergences for $h^{-1}+1 < 24$. Using the fitted constants $\tilde{K}_0$ and K_1, the optimal fidelity $h^* \in \{h_1, \ldots, h_{15}\}$ is found using a brute-force search with Figure 2.7 showing the optimal number of grid points $(h^*)^{-1} + 1$ for different tolerances ϵ on the MSE. We see that once the tolerance ϵ becomes less than 10^{-2} the optimal number of grid points levels off as there are diminishing returns for investing more computational resources in the online phase. Note that this is in contrast to the theoretical analysis of Section 2.3.4 where we showed that the fidelity $h^* \to 0$ as $\epsilon \to 0$ and is because here we do not consider any surrogate models with $h^{-1} + 1$ between 64 and 256 grid points. Figure 2.8 shows the theoretical total costs and tolerance ϵ trade-off for both the CAIS estimator (2.39) as well as the MFIS estimator (2.12) with a fixed fidelity. Again we consider the fixed low-fidelity model to be h_3 with $h_3^{-1} + 1 = 16$ grid points. For the true MSE estimated in the right plot of Figure 2.8, we first estimated a ground truth reference value $\hat{f}$ of $\mathbb{E}_\pi[f]$ using $\hat{f}_{h_0,10^5}$ averaged over $N_2 = 100$ independent trials using equation (2.56). Then the MSE is estimated by averaging over $N_3 = 2500$ independent trials using equation (2.57). We again see that for error tolerances as small as 10^{-2} the CAIS estimator enjoys an order of magnitude speedup over the fixed fidelity MFIS estimator with

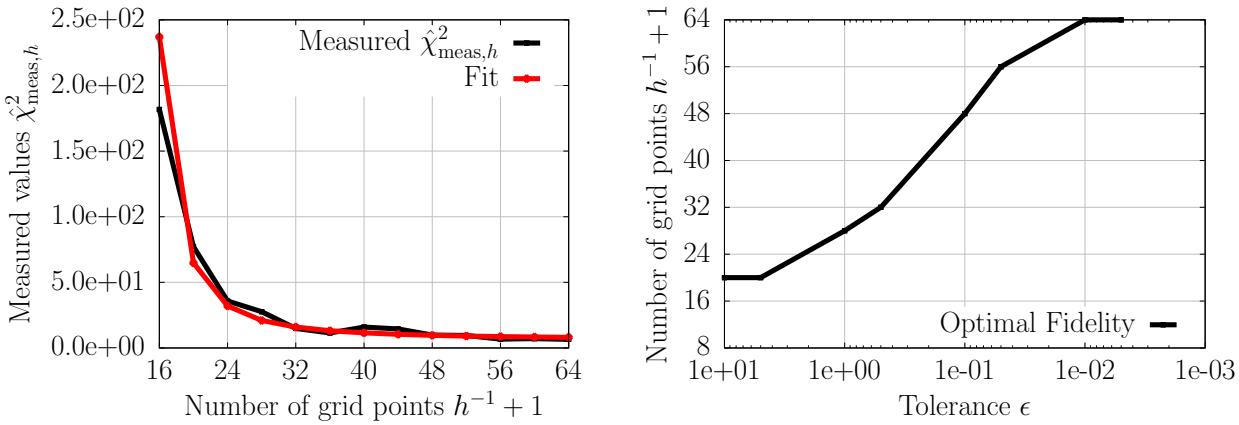


Figure 2.7: (**Left**) The measured χ^2 divergences $\hat{\chi}^2_{\text{meas},h}$ given by (2.54) of $\chi^2(\pi \,||\, \mu_{h_i})$ for $i = 1, \ldots, 15$ and the corresponding fitted curve. (**Right**) The selected fidelity $h^* \in \{h_1, \ldots, h_{15}\}$ for the optimal number of elements $(h^*)^{-1}$ from the optimization (2.34) as the tolerance ϵ on the MSE changes.

the high-fidelity density π used to fit the biasing density. As the tolerance $\epsilon \to 0$ shrinks both the costs of the context-aware and high-fidelity alone estimators are asymptotically the same and both outperform the the low-fidelity alone MFIS estimator due to the poor biasing density that is constructed.

2.5.3 Advection-diffusion Problem

In this example, we consider a concentration of gas in air that diffuses throughout a domain with advection given by wind. In particular, we want to infer the initial center of the concentration of gas given observations of the concentration at a later time. The forward model is taken from hIPPYlib [2] [Villa et al., 2016, Villa et al., 2018, Villa et al., 2021] with minor modifications.

[2] `https://hippylib.github.io/tutorials_v3.0.0/4_AdvectionDiffusionBayesian/`

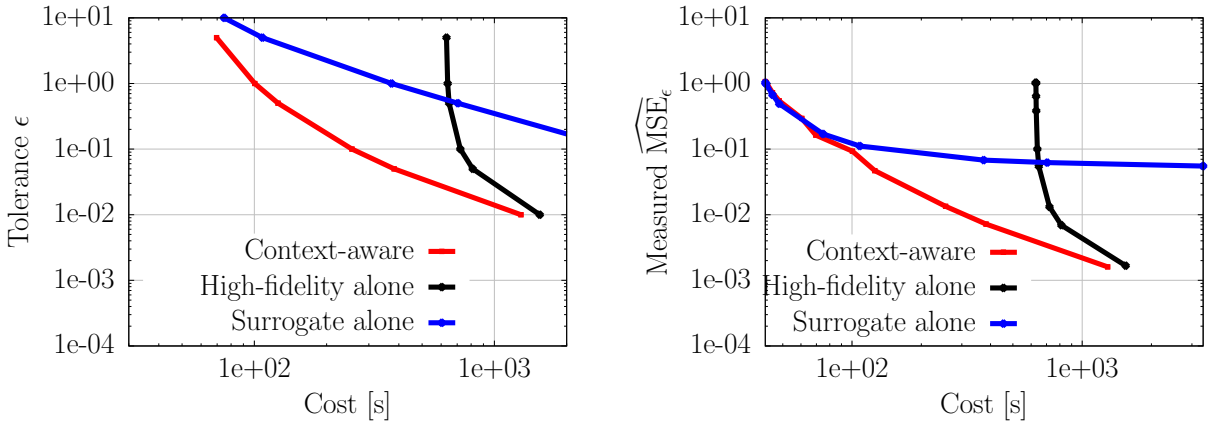


Figure 2.8: (**Left**) The error tolerance ϵ against the total costs (2.33) (runtime in seconds) to fit the Laplace approximation μ_{h^*} of $\pi^{(h^*)}$ and compute the CAIS estimator $\hat{f}^{\text{CAIS}}_{h^*,N^*}$ with N^* samples compared against the MFIS estimator (2.12) with a fixed fidelity. (**Right**) The estimated MSE, $\widehat{\text{MSE}}_\epsilon$, with respect to ground truth reference value against the total costs.

Problem Setup

Following the setup in hIPPYlib, consider the domain

$$\Omega = [0,1]^2 \setminus ([0.25, 0.5] \times [0.15, 0.4] \cup [0.6, 0.75] \times [0.6, 0.85]) \subset \mathbb{R}^2 \,,$$

as the unit square with two rectangular holes. For this problem the parameter corresponds to a point in the domain, and so we set $\Theta = \Omega$. Now let $u : \Omega \times [0,1] \times \Theta \to \mathbb{R}$ denote the concentration of the gas at position $\mathbf{x} \in \Omega$ and time $t \in [0,1]$ which is governed by the following advection-diffusion PDE

$$\begin{aligned} \partial_t u(\mathbf{x}, t; \boldsymbol{\theta}) - \kappa \Delta u(\mathbf{x}, t; \boldsymbol{\theta}) + \mathbf{v}(\mathbf{x}) \cdot \nabla u(\mathbf{x}, t; \boldsymbol{\theta}) &= 0, \quad (\mathbf{x}, t) \in \Omega \times [0,1] \,, \\ u(\mathbf{x}, 0; \boldsymbol{\theta}) &= \mathrm{e}^{-10(x_1-\theta_1)^2 - 10(x_2-\theta_2)^2}, \quad \mathbf{x} \in \Omega \,, \\ \kappa \nabla u(\mathbf{x}, t; \boldsymbol{\theta}) &= \boldsymbol{n}, \quad (\mathbf{x}, t) \in \partial\Omega \times [0,1] \,, \end{aligned} \tag{2.59}$$

where the initial concentration is parameterized to be a Gaussian centered at $\boldsymbol{\theta} \in \Theta$. Here $\kappa = 10^{-3}$ is the diffusion coefficient that controls the rate at which the gas diffuses, $\boldsymbol{n}$ is the

outward unit normal vector from the boundary, and $\mathbf{v} : \Omega \to \mathbb{R}^2$ is the velocity field for the wind that moves the concentration around the domain. In particular, the velocity field $\mathbf{v}$ is the solution of the steady-state Navier-Stokes equation

$$\begin{aligned} -\frac{1}{\mathrm{Re}}\mathbf{v}(\mathbf{x}) + \nabla q(\mathbf{x}) + \mathbf{v}(\mathbf{x}) \cdot \nabla \mathbf{v}(\mathbf{x}) = \mathbf{0}, \quad & \mathbf{x} \in \Omega\,, \\ \nabla \cdot \mathbf{v}(\mathbf{x}) = 0, \quad & \mathbf{x} \in \Omega\,, \\ \mathbf{v}(\mathbf{x}) = \mathbf{g}(\mathbf{x}), \quad & \mathbf{x} \in \partial\Omega\,, \end{aligned} \tag{2.60}$$

with the left and right walls driving the flow through the boundary condition that $\mathbf{v}(\mathbf{x}) = \mathbf{g}(\mathbf{x})$ (see Figure 2.9). In Equation (2.60) the constant $\mathrm{Re} = 10^2$ is the Reynold's number of the surrounding air, $q : \Omega \to \mathbb{R}$ is the corresponding pressure field, and external force $\mathbf{g} : \partial\Omega \to \mathbb{R}$ given by

$$\mathbf{g}(\mathbf{x}) = \begin{cases} \boldsymbol{e}_2\,, & x_1 = 0 \\ -\boldsymbol{e}_2\,, & x_1 = 1 \\ \mathbf{0}\,, & x_1 \in (0,1)\,, \end{cases}$$

where $\boldsymbol{e}_2 = (0,1)^\top$. From the left plot in Figure 2.9 we see that indeed the the wind is flowing towards the top of the domain along the left wall where $x_1 = 0$ and towards the bottom of the domain along the right wall where $x_1 = 1$. The dependence on the solution $u(\mathbf{x}, t; \boldsymbol{\theta})$ of (2.59) on the parameter $\boldsymbol{\theta}$ is only through the initial condition $u(\mathbf{x}, 0; \boldsymbol{\theta})$ and that, in particular, the velocity field $\mathbf{v}$ for the wind (2.60) is independent of $\boldsymbol{\theta}$. Thus, we can precompute $\mathbf{v}$ once (for each mesh). Although the PDE (2.59) is linear for u, it is nonlinear in the parameters $\boldsymbol{\theta}$ due to the form of the initial concentration. To be consistent with the other numerical examples 2.5.1 and 2.5.2, let the interpolation operator $\mathcal{I}^{\mathrm{int}} : \Theta \to C^2[0,1]$ be defined by

$$\mathcal{I}^{\mathrm{int}}(\boldsymbol{\theta}) = \mathrm{e}^{-10(x_1-\theta_1)^2 - 10(x_2-\theta_2)^2}\,,$$

which maps the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$ to the initial concentration field. The forward models $F^{(h)} : C^2[0,1] \to C[0,1]$ approximate the solution operator which maps the initial concentration $u(\mathbf{x}, 0; \boldsymbol{\theta})$ to the final concentration $u(\mathbf{x}, 1; \boldsymbol{\theta})$ at time $t = 1$. Following [Villa et al., 2016, Villa et al., 2018, Villa et al., 2021], the spatial domain Ω is discretized with first order Lagrange finite elements and then integrates in with implicit Euler in time to obtain the approximation $u^{(h)}(\mathbf{x}, 1; \boldsymbol{\theta})$. Here the fidelity h is the maximum width of a cell in the mesh and decreases as the number of cells, and hence degrees of freedom, increases. For the high-fidelity approximation F to the solution operator, a discretization with 14,313 degrees of freedom and a time step size of 10^{-3} is used. The surrogate models using the approximations $F^{(h)}$ are defined similarly with the total number of degrees of freedom in the discretized system ranging from 20 to 3,661 and with a time step size of 10^{-2}. For each surrogate model $F^{(h)}$ and high-fidelity model F, the computation of the velocity field $\mathbf{v}$ is done by solving the steady-sate Navier-Stokes equation (2.60) by discretizing with mixed quadratic and linear finite elements for the velocity and pressure components, respectively, on the same spatial mesh used to solve (2.59) for the concentration u. The resulting nonlinear system is then solved using Newton's method until the relative norm of the gradient is below 10^{-4} its initial value. Pointwise observations of the concentration at the final time are taken at four points in the bottom-right quadrant of the domain as in Figure 2.9 with observation operator $\mathcal{B}^{\text{obs}} : C[0,1] \to \mathbb{R}^4$ defined by

$$\mathcal{B}^{\text{obs}}(u^{(h)}) = \begin{pmatrix} u^{(h)}(\mathbf{x}_1, 1; \boldsymbol{\theta}) \\ u^{(h)}(\mathbf{x}_2, 1; \boldsymbol{\theta}) \\ u^{(h)}(\mathbf{x}_3, 1; \boldsymbol{\theta}) \\ u^{(h)}(\mathbf{x}_4, 1; \boldsymbol{\theta}) \end{pmatrix}, \quad \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4 \end{pmatrix} = \begin{pmatrix} 2/3 & 5/6 & 2/3 & 5/6 \\ 1/6 & 1/6 & 1/3 & 1/3 \end{pmatrix}.$$

The full parameter-to-observable maps $G^{(h)} : \mathbb{R}^2 \to \mathbb{R}^4$ are given by $G^{(h)} = \mathcal{B}^{\text{obs}} \circ F^{(h)} \circ \mathcal{I}^{\text{int}}$. Because the parameter domain $\Theta = \Omega$ is bounded, we do not have to worry about the biasing

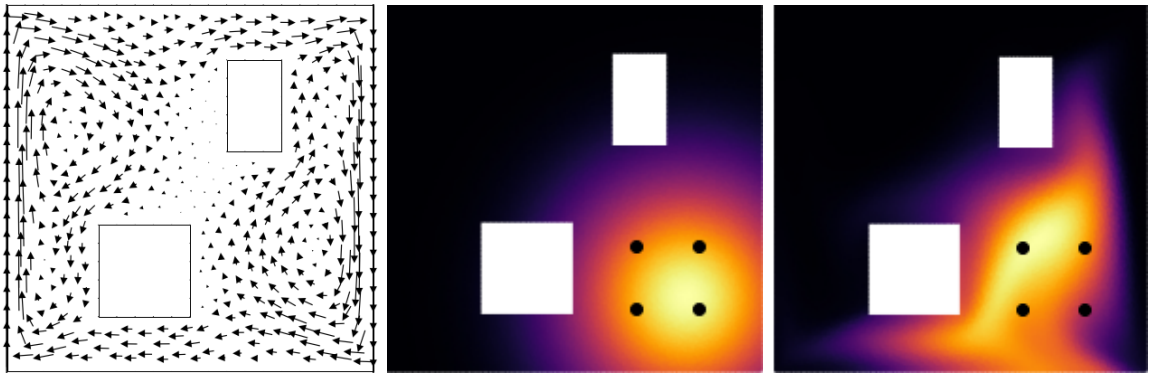


Figure 2.9: **(Left)** The velocity field $\mathbf{v}$ throughout the domain Ω with the two rectangular barriers. **(Middle)** The true initial concentration $u(\mathbf{x}, 0; \boldsymbol{\theta}^\star)$ centered at $\boldsymbol{\theta}^\star = (0.8, 0.2)^\top \in \Theta$. **(Right)** The true concentration $u(\mathbf{x}, 1; \boldsymbol{\theta}^\star)$ at the final time $t = 1$ when we observe the solution. In both middle and right plots the observation points are shown as the four black dots in the bottom-right corner.

density or surrogate densities decaying too fast for importance sampling to fail as long as they are absolutely continuous with respect to each other. In particular, Assumptions 2, 3, and 4 are satisfied. Moreover, Assumptions 6 and 5 are satisfied because the forward models F and $F^{(h)}$, and hence the parameter-to-observable maps G and $G^{(h)}$, are differentiable on a compact domain and therefore globally Lipschitz. As a result Theorem 3 may be applied to obtain the cost complexity of the CAIS estimator (2.39) given by Theorem 2.

Setup of the inverse problem

A single observation

$$\mathbf{y} = G(\boldsymbol{\theta}^\star) + \boldsymbol{\eta}\,, \quad \boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{\Gamma})\,,$$

is generated from the true initial concentration center $\boldsymbol{\theta}^\star = (0.8, 0.2)^\top$ and perturbed by Gaussian noise with covariance $\boldsymbol{\Gamma} = 8.876 \times 10^{-3} \mathbf{I}_{4\times 4}$. The standard deviation $\sqrt{8.876 \times 10^{-3}}$ of the added noise is 10% of the norm of the true solution $u(\mathbf{x}, 1; \boldsymbol{\theta}^\star)$. For the prior we take a Gaussian with mean $\boldsymbol{\mu}_0 = (0.75,\ 0.25)^\top$ and covariance $\boldsymbol{\Sigma}_0 = 10^{-2} \mathbf{I}_{2\times 2}$ and restrict it to have zero density outside of the domain Ω. The restriction of the prior to be supported over the domain Θ is only necessary to ensure that the posterior has zero density outside of the

domain. Practically this is straightforward to implement because samples can be drawn by sampling from the unrestricted distribution $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and rejecting those that lie outside of Ω. Moreover, the prior is only needed up to a normalizing constant since self-normalized importance sampling (2.4) is used.

Fitting the Laplace approximation

Similar to pilot studies of the previous two examples, we fit the Laplace approximation to each surrogate posterior $\pi^{(h)}$ using Newton's method. However, instead of using finite differences to approximate the gradients and Hessians of the potentials $\Phi^{(h)}$, here we leverage fast adjoint solvers available in hIPPYlib to efficiently compute the derivatives. In particular, hIPPYlib computes the gradient and Hessian of the negative log-likelihood

$$\frac{1}{2}\left\|\mathbf{y} - \mathcal{B}^{\text{obs}} \circ F^{(h)}(u_0)\right\|^2_{\boldsymbol{\Gamma}^{-1}}$$

with respect to the initial concentration u_0. To obtain the derivatives with respect to the parameters $\boldsymbol{\theta}$ we apply the chain rule with the derivatives of the interpolation operator with respect to $\boldsymbol{\theta}$. Although (2.59) is a linear PDE and hence the solution operator $F^{(h)}$ is linear, the interpolation operator $\mathcal{I}^{\text{int}}$ is nonlinear and thus the Hessian must be re-evaluated at every step of Newton's method. In particular, evaluating the gradient requires solving both the forward problem as well as the adjoint problem while the Hessian requires two additional linear solves for the forward incremental and adjoint incremental equations. We consider the cost of each of these four linear solves to be equivalent and therefore set the number of surrogate model evaluations to be 4 per iteration of Newton's method. To fit the Laplace approximation we start at a random initial point in the bottom-right quadrant $\boldsymbol{\theta}_0 \in [0.5,\ 1] \times [0, 0.5]$ and find that 10 Newton iterations, but not fewer, is sufficient for the norm of the gradient of the log posterior to achieve machine precision. Thus, the total

number of offline evaluations of the surrogate model $G^{(h)}$ is set to $N_0 = 40$.

Pilot study

To estimate the constants $\tilde{K}_0, K_1$, we follow the same procedure as in the steady-state heat flow and Euler Bernoulli problems with the exception that here we fit the Laplace approximations using the adjoint method for the derivatives. We estimate the chi-squared divergences using (2.53) with 5×10^5 total samples and then fit a curve of the form $\tilde{K}_0 \mathrm{e}^{K_1 \delta(h)}$. The resulting fitted curve as well as the estimated χ^2 values are shown in Figure 2.10, where we see the same behavior as in the earlier examples of the χ^2 divergence quickly leveling off. Recall that the fidelity h corresponds to the max-width of the mesh in the spatial domain. For the surrogate error $\delta(h)$ we fit a curve of the form $\delta(h) \propto h^\alpha$, following from the convergence theory of finite elements. To do this we first compute a reference solution $u^{(h_0)}$ for a small max mesh width h_0 and then compare the L^2 error

$$\left\| u^{(h_0)} - u^{(h)} \right\|$$

of the surrogate model solutions $u^{(h)}$. Similarly, to obtain the cost function $c(h)$, we measure the runtime of each surrogate and high fidelity model and average over 10,000 trials. Using the measured costs we fit a curve of the form $c(h) = c_0 h^{-\beta}$ since the logarithm of the maximum mesh width scales proportionally to the logarithm of the number of degrees of freedom, for the particular mesh obtained with FEniCS [Alnaes et al., 2015].

Results

Using the constants $\tilde{K}_0, K_1$ as well as the functions c, δ obtained during the pilot study, we can formulate the optimization problem (2.34) for the optimal fidelity h^* required for the CAIS estimator. Although computing the Laplace approximation in the offline phase must

be done in serial, the online phase is trivially parallelizable as the samples may be drawn and re-weighted according to the high-fidelity posterior density π independently. Thus, when considering total runtime as the costs we may scale the necessary number of samples by the number of available processors. Here we paralleilize the online phase over $n_{\text{proc}} = 64$ processors. The optimal fidelity or max mesh width in this case is given by the solution to the optimization problem

$$\min_{h>0} \quad \frac{4 \|f\|_{L^\infty}^2 \tilde{K}_0 C^{\text{high}}}{\epsilon n_{\text{proc}}} \mathrm{e}^{K_1 \delta(h)} + N_0 c(h) \,. \tag{2.61}$$

Since here we consider only finitely many surrogate models we do a brute-force search to solve (2.61) for the optimal mesh width, which in turn gives the optimal number of degrees of freedom for the discretization as shown in Figure 2.10. We again see that optimal number of degrees of freedom levels off after shrinking the tolerance ϵ below 10^{-2} since the surrogate model is already sufficiently accurate. Figure 2.11 shows the theoretical speedup in runtime predicted by the optimization problem (2.61) versus the actual measured speedup for the CAIS estimator (2.39) over the fixed high and low fidelity MFIS estimators. The ground truth reference value used to compute the MSE was estimated using the MFIS estimator (2.12) with the high-fidelity surrogate model and 10^6 total samples. In the regime where the tolerance ϵ is moderate, e.g. larger than 10^{-2}, the CAIS estimator (2.39) is able to take advantage of the cheap surrogate models to attain large speedups over the fixed high-fidelity estimator. However, as $\epsilon \to 0$ this speedup over the MFIS estimator with the high-fidelity surrogate model diminishes as optimal fidelity shown in Figure 2.10 levels off. Alternatively, as more computational resources are invested into the CAIS estimator, a more accurate surrogate model is required eliminating speedup in the offline phase. When compared to the MFIS estimator that only uses a low fidelity surrogate model we observe the opposite, which is that for moderate error tolerances an inaccurate surrogate model with $\chi^2(\pi \,||\, \mu_h)$ large is still

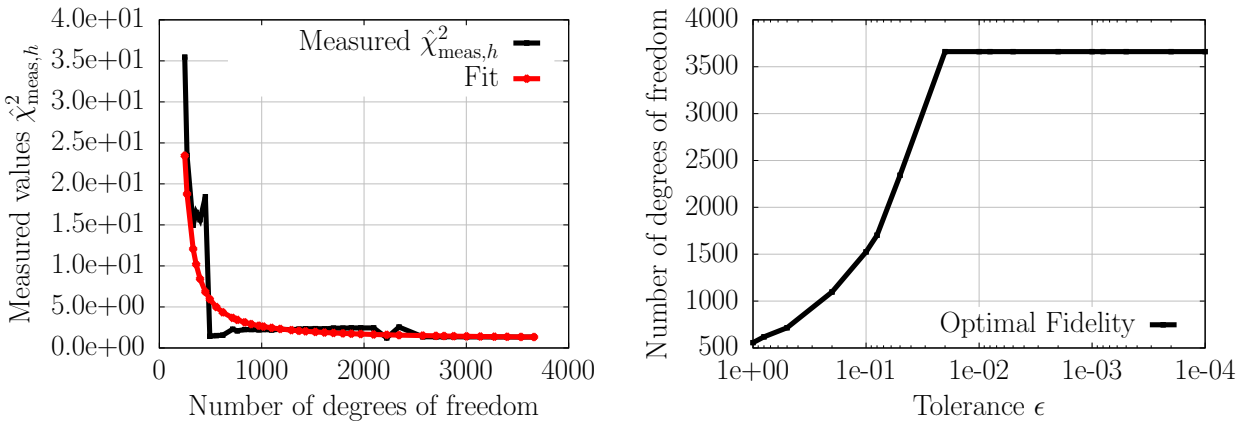


Figure 2.10: **(Left)** The measured and fitted values of $\chi^2(\pi \mid\mid \mu_h) + 1$ for each Laplace approximation to a surrogate model $\pi^{(h)}$. **(Right)** The optimal number of degrees of freedom corresponding the optimal mesh width as given by (2.61) for different tolerances ϵ.

sufficient, As $\epsilon \to 0$ and the online costs begin to dominate, larger effective sample sizes are required giving rise to speedup over low-fidelity only estimators that derived poor biasing densities. Note that when comparing the left and right plots of Figure 2.11, the plot on the left shows the tolerance ϵ which is an upper bound on the MSE through (2.8). Moreover, the discrepancy between the MSE and the tolerance ϵ can be accounted for with the fact that the bound (2.8) is uniform over all test functions and that for a fixed test function f one may obtain a tighter bound. Despite this, the curves in both the left and right plots behave similar qualitatively i.e. show similar speedups and asymptotic cost complexity rates.

Extension to 12-dimensional parameter

So far the three examples we have looked at are in relatively low parameter dimensions (6, 6, and 2, respectively). For the advection-diffusion problem we can easily change the dimension of the parameter to correspond to additional centers for the initial concentration. In particular, we now look at a 12-dimensional extension of the advection-diffusion problem by considering six unique centers for the initial concentration (2 dimensions per center). The forward models F and $F^{(h)}$, which approximate the solution operator, remain the same, but

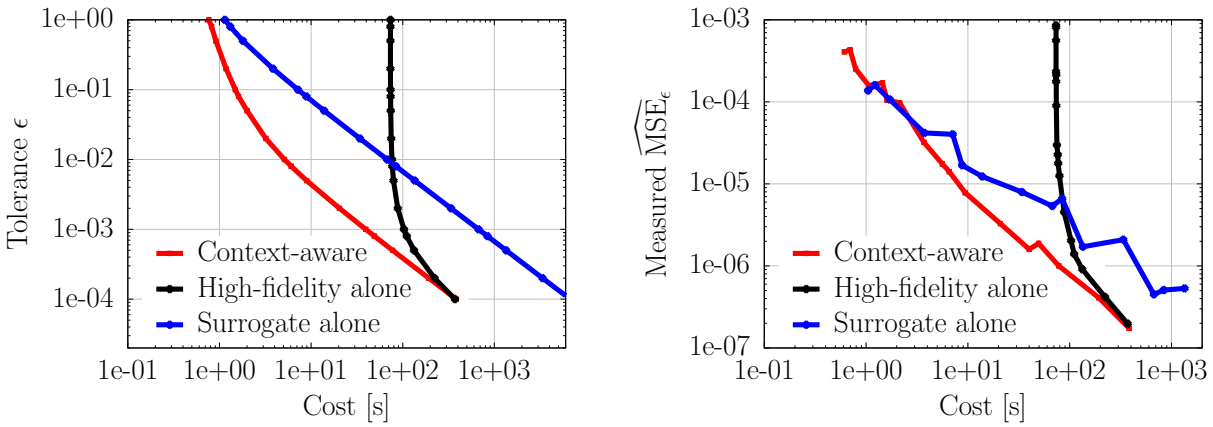


Figure 2.11: **(Left)** The tolerance ϵ on the MSE vs. the theoretical cost in seconds of runtime of the entire computational procedure. **(Right)** The estimated MSE of the CAIS estimator with the optimal surrogate model vs. cost. For the CAIS estimator, the MSE is computed by averaging over 100 independent trials. For the fixed high and low-fidelity estimators the MSE is estimated by averaging over 50 independent trials.

we now define the interpolation operator to be

$$\mathcal{I}^{\text{int}}(\boldsymbol{\theta}) = \sum_{i=1}^{6} \mathrm{e}^{-10(x_1-\theta_{2i-1})^2-10(x_2-\theta_{2i})^2}, \quad \mathbf{x} \in \Omega\,,$$

which maps $\boldsymbol{\theta} \in \mathbb{R}^{12}$ to the initial concentration $u(\mathbf{x}, 0; \boldsymbol{\theta}) \in C^2(\Omega)$. We also add four new observation points, for eight total, in the top left corner of the domain Ω, which are the same as the observation points shown in 2.9 but reflected across the line $x_1 = x_2$. Therefore, the observation operator $\mathcal{B}^{\text{obs}} : C^2[0,1] \to \mathbb{R}^8$ and hence $G : \mathbb{R}^{12} \to \mathbb{R}^8$ as well as the surrogate models $G^{(h)}$. To generate the synthetic data $\mathbf{y}$ we also increase the standard deviation of the added noise $\boldsymbol{\eta}$ to 20% of the norm of the true solution $\|u(\cdot, 1; \boldsymbol{\theta}^\star)\|_{L^\infty}$. We set the prior π_0 to be Gaussian with a covariance of $\boldsymbol{\Sigma}_0 = 4 \times 10^{-3}\mathbf{I}_{12\times 12}$. The rest of the problem set up is the same as the 2-dimensional problem as well as performing an analogous pilot study. Figures 2.12 and 2.13 show that the context-aware estimator outperforms both the estimator that uses the high-fidelity alone and the estimator that only uses the low-fidelity model for constructing the biasing density. This example also suggests that the CAIS and MFIS

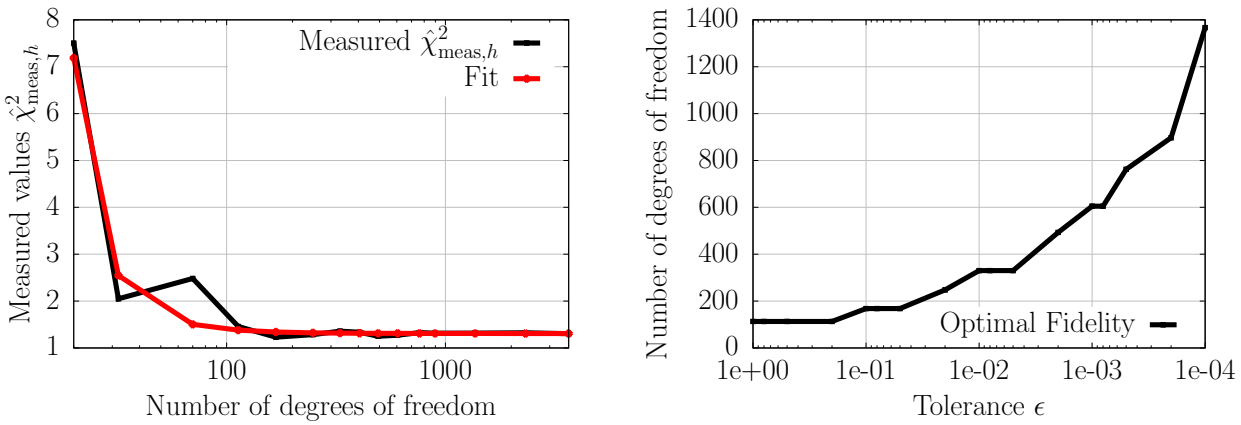


Figure 2.12: **(Left)** The estimated values of $\chi^2(\pi \,||\, \mu_h) + 1$ for each surrogate model $\pi^{(h)}$ as well as the fitted curve with constants $\tilde{K}_0$ and K_1. **(Right)** The optimal number of degrees of freedom as given by (2.61) for different tolerances ϵ.

estimators are independent of the dimension, a notable feature of Monte Carlo methods, as long as χ^2 divergence does not blow up with the dimension d. This means that the CAIS can achieve speedup over a fixed fidelity MFIS estimator as long as the surrogate models and biasing densities are sufficiently accurate to provide a useful approximation with small χ^2 divergence.

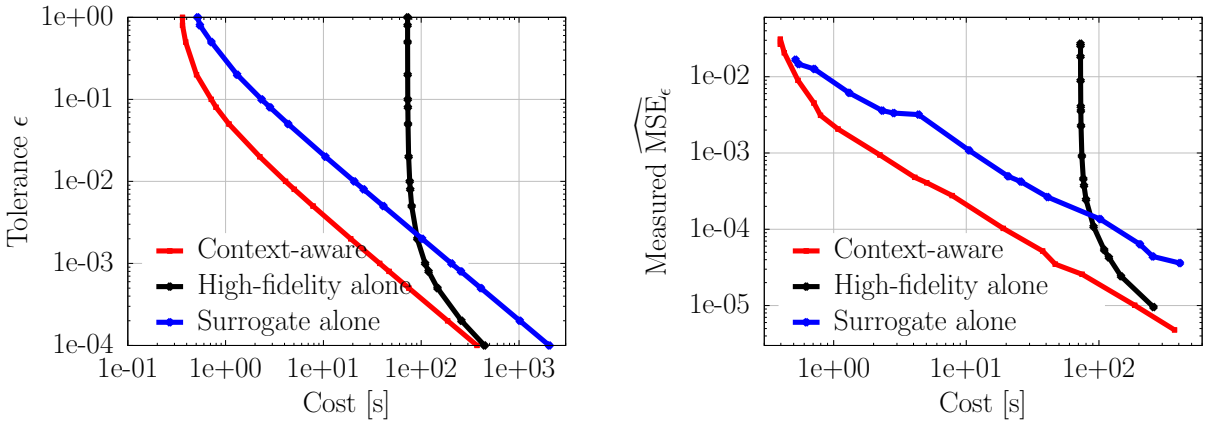


Figure 2.13: **(Left)** The tolerance ϵ on the MSE vs. the theoretical cost in seconds of runtime of the entire computational procedure. **(Right)** The estimated MSE of the CAIS estimator with the optimal surrogate model vs. cost. For the CAIS estimator, the MSE is computed by averaging over 100 independent trials. For the fixed high and low-fidelity estimators the MSE is estimated by averaging over 50 independent trials.

Chapter 3

Multilevel Stein variational gradient descent

In this chapter we present the published work [Alsup et al., 2021] as well as our pre-print [Alsup et al., 2022]. Many computational applications admit a hierarchy of available surrogate models that become increasingly accurate and expensive as the level increases. Here we consider a multilevel extension to Stein variational gradient descent (SVGD), that leverages such a hierarchy of surrogate models to learn increasingly accurate densities. First proposed by Liu and Wang [Liu and Wang, 2016], SVGD mimics traditional variational inference but avoids an explicit parametrization of the approximating distribution by instead updating an ensemble of particles [Liu et al., 2019]. More efficient extensions of SVGD seek to exploit curvature of the target distribution such as in the Stein variational Newton method [Detommaso et al., 2018] or by using adaptive kernels [Duncan et al., 2019, Wang et al., 2019]. SVGD has also been extended to take advantage of low-dimensional structure in the target distribution [Chen et al., 2019] as well as having a gradient-free version [Han and Liu, 2018] that instead computes gradients with respect to a surrogate model and then re-weights using

importance sampling. Other variants, which are more akin to traditional MCMC methods based on [Ma et al., 2015], include stochastic analogues [Gallego and Insua, 2020, Leviyev et al., 2022] that are highly efficient and can be Metropolized to eliminate bias at each iteration. Most of the analysis of SVGD stems from the works [Liu, 2017, Lu et al., 2019] which showed that SVGD follows a gradient flow with respect to the KL divergence in the mean-field and continuous-time limit. The work [Chewi et al., 2020] showed an analogous result for the chi-squared divergence. Limited progress, primarily in the work [Korba et al., 2020], has been made towards extending this analysis to obtain pre-asymptotic and finite particle convergence results and largely remains an open problem. Finally, other directions of analysis include understanding the performance of SVGD in high-dimensions [Ba et al., 2019].

3.1 Stein variational gradient descent

SVGD [Liu and Wang, 2016] is a general purpose inference method that performs variational inference in a nonparametric fashion using an ensemble of particles. Because SVGD is nonparametric, it does not require specifying a parametrized family of distributions to be optimized over as is the case with fitting transport map approximations, normalizing flows, or Gaussian mixture models. Similar to other variational approaches for inference, SVGD seeks to find an approximation μ that minimizes the KL divergence $\mathrm{KL}\left(\mu \,||\, \pi\right)$ to the target density π. While more traditional variational approaches attempt to solve

$$\mu^\star = \min_{\mu \in \mathcal{P}} \quad \mathrm{KL}\left(\mu \,||\, \pi\right) = \int_{\mathbb{R}} \mu(\boldsymbol{\theta}) \log \frac{\mu(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}$$

directly over a family of densities $\mathcal{P}$, SVGD uses an iterative approach by performing steepest descent to update the density μ. For SVGD, one starts with an initial density μ_0 and updates

it in the steepest descent direction of a transport map that minimizes the KL divergence. Instead of minimizing over a parametrization of maps, such as transport map or normalizing flow approximations, SVGD minimizes over maps in a reproducing kernel Hilbert space (RKHS). Let $\mathcal{H}$ be a RKHS with positive definite kernel $K : \Theta \times \Theta \to \mathbb{R}$ of functions $g : \Theta \to \mathbb{R}$ so that

$$\langle K(\boldsymbol{\theta}, \cdot),\ g(\cdot) \rangle_{\mathcal{H}} = g(\boldsymbol{\theta}), \quad \forall g \in \mathcal{H},\ \boldsymbol{\theta} \in \Theta\,,$$

and let $\mathcal{H}^d \simeq \mathcal{H} \times \cdots \times \mathcal{H}$ be the corresponding RKHS of vector fields $\mathbf{g} = (g_1, \ldots, g_d) : \Theta \to \mathbb{R}^d$. For any density μ such that μ is absolutely continuous with respect to π and

$$\mathrm{KL}\left(\mu \,||\, \pi\right) < \infty\,,$$

define the KL functional $J_\mu : \mathcal{H}^d \to \mathbb{R}$ by

$$J_\mu(\mathbf{g}) = \mathrm{KL}\left((\mathbf{I} - \mathbf{g})_{\#}\mu \,||\, \pi\right)\,,$$

where $\mathbf{g} \in \mathcal{H}^d$. The functional $J_\mu(\mathbf{g})$ evaluates the KL divergence of the updated density $(\mathbf{I} - \mathbf{g})_{\#}\mu$ to the target density π, and so the functional gradient at zero is the function $\nabla_\mu J(\mathbf{0}) \in \mathcal{H}^d$ such that

$$\langle \nabla J_\mu(\mathbf{0}),\ \mathbf{g} \rangle_{\mathcal{H}^d} = \lim_{\epsilon \to 0} \frac{\mathrm{KL}\left((\mathbf{I} - \epsilon\mathbf{g})_{\#}\mu \,||\, \pi\right) - \mathrm{KL}\left(\mu \,||\, \pi\right)}{\epsilon}\,.$$

The initial SVGD work [Liu and Wang, 2016] showed that the functional gradient $\nabla J_\mu(\mathbf{0})$ has a closed form solution using the RKHS structure of $\mathcal{H}$ and Stein's identity. In particular, it was shown that

$$\nabla J_\mu(\mathbf{0})\left(\boldsymbol{\theta}\right) = -\mathbb{E}_{\mathbf{z} \sim \mu}\left[K(\mathbf{z}, \boldsymbol{\theta}) \nabla \log \pi(\mathbf{z}) + \nabla_1 K(\mathbf{z}, \boldsymbol{\theta})\right]\,, \tag{3.1}$$

where ∇_1 denotes the gradient with respect to the first argument only.

Continuous-time SVGD

From a particle's perspective, SVGD starts with an initial particle $\boldsymbol{\theta}_0 \sim \mu_0$ and evolves it along the steepest descent direction, also known as the mean-field characteristic flow [Lu et al., 2019],

$$\dot{\boldsymbol{\theta}}_t = -\nabla J_{\mu_t}(\mathbf{0})(\boldsymbol{\theta}_t)\,, \tag{3.2}$$

where μ_t denotes the density of $\boldsymbol{\theta}_t$ at time $t \geq 0$. At any instance in time t the particle is moving in the direction $-\nabla J_{\mu_t}(\mathbf{0})$ which depends on the current density of the particle μ_t. This density is governed by the following nonlinear Fokker-Planck equation

$$\begin{aligned} \partial_t \mu_t(\boldsymbol{\theta}) &= \nabla \cdot (\mu_t(\boldsymbol{\theta}) \nabla J_{\mu_t}(\mathbf{0})(\boldsymbol{\theta})) \\ &= -\nabla \cdot (\mu_t(\boldsymbol{\theta}) \mathbb{E}_{\mathbf{z}\sim\mu_t} \left[K(\mathbf{z}, \boldsymbol{\theta}) \nabla \log \pi(\mathbf{z}) + \nabla_1 K(\mathbf{z}, \boldsymbol{\theta})\right])\,. \end{aligned} \tag{3.3}$$

We refer to the approximation μ_t that solves (3.3) as the *continuous-time* SVGD approximation in contrast to a discrete-time version which we present in the next section. An important result concerned with the convergence of the solution μ_t to (3.3) derived in [Liu, 2017, Theorem 3.4] is that, for the solution μ_t of (3.3), it holds that

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{KL}\left(\mu_t \,||\, \pi\right) = -\mathbb{D}(\mu_t \,||\, \pi)^2\,, \tag{3.4}$$

where

$$\mathbb{D}(\mu_t \,||\, \pi) = \max_{\mathbf{g}\in\mathcal{H}^d} \left\{\mathbb{E}_{\boldsymbol{\theta}\sim\mu_t}[\nabla \log \pi(\boldsymbol{\theta})^\top \mathbf{g}(\boldsymbol{\theta}) + \nabla \cdot \mathbf{g}(\boldsymbol{\theta})] : \|\mathbf{g}\|_{\mathcal{H}^d} \leq 1\right\}$$

is the Stein discrepancy. Equation (3.4) guarantees that the solution μ_t to (3.3) decreases the KL but does not necessarily need to converge to 0. The solution μ_t may fail to converge to π whenever the space $\mathcal{H}$ is not sufficiently rich and so the kernel K must be chosen

appropriately. We note that the converse is true however, $\mathbb{D}(\mu \,||\, \pi) = 0$ if $\mu = \pi$.

Discrete-time SVGD

Continuous-time SVGD requires integrating the characteristic flow (3.2) exactly to obtain a sample $\boldsymbol{\theta}_t \sim \mu_t$ where μ_t solves (3.3). However, this may be difficult because the flow (3.2) depends on the current density μ_t as well as potentially no closed-form solution for μ_t existing. Instead we may apply the explicit Euler method to integrate the flow (3.2) to obtain

$$\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \delta \mathbf{g}_\tau(\boldsymbol{\theta}_\tau)\,, \qquad \mathbf{g}_\tau = \nabla J_{\mu_\tau}(\mathbf{0})\,, \tag{3.5}$$

where $\tau = 1, 2, \ldots$ now corresponds to an iteration and $\delta > 0$ is a predetermined step size. The updated density $\mu_{\tau+1}$ becomes the pushforward of the previous density

$$\mu_{\tau+1} = (\mathbf{I} - \delta \mathbf{g}_\tau)_{\#}\mu_\tau\,, \quad \tau = 1, 2, \ldots. \tag{3.6}$$

We refer to the density μ_τ obtained by the sequence of updates (3.6) as the *discrete-time* SVGD approximation.

Remark 8. Both the continuous-time (3.3) and discrete-time (3.6) SVGD approximations assume that the gradients $\nabla_\mu J(\mathbf{0})$ are evaluated exactly, which may not be achievable in practice and gives rise to the particle formulation presented in Section 3.5.

A bound similar to (3.4) was shown in [Liu, 2017, Theorem 3.3] to hold for discrete-time SVGD as long as the step size δ is sufficiently small

$$\mathrm{KL}\left(\mu_{\tau+1} \,||\, \pi\right) \leq \mathrm{KL}\left(\mu_\tau \,||\, \pi\right) - \delta(1 - \delta B)\mathbb{D}(\mu_\tau \,||\, \pi)^2\,, \tag{3.7}$$

where the constant B depends on the kernel K and the target density π. As with continuous-time SVGD, the sequence of densities $(\mu_\tau)_{\tau \geq 1}$ minimizes the KL divergence.

3.2 Continuous-time single-level SVGD and MLSVGD

In this section we consider only continuous-time SVGD and MLSVGD and reserve discussion of their discrete-time counterparts to Section 3.3. SVGD requires being able to evaluate the score function $\nabla \log \pi$ exactly to ensure that the solution μ_t to (3.3) converges to π. However, in the setting where the score function is intractable, which is often the case for Bayesian inverse problems, we may still obtain an approximation μ to arbitrary accuracy ϵ in the Hellinger metric [Tsybakov, 2009] defined as

$$\mathrm{d}_{\mathrm{Hell}}\left(\pi,\ \mu\right)^2 = \frac{1}{2}\int_{\Theta}\left(\sqrt{\pi(\boldsymbol{\theta})} - \sqrt{\mu(\boldsymbol{\theta})}\right)^2 \mathrm{d}\boldsymbol{\theta}\,, \tag{3.8}$$

so that

$$\mathrm{d}_{\mathrm{Hell}}\left(\mu,\ \pi\right) \leq \epsilon$$

given that sufficiently accurate surrogate models are available. Let $(\pi^{(\ell)})_{\ell=1}^{\infty}$ denote a sequence of approximating densities over $\Theta \subset \mathbb{R}^d$ and integrating continuous-time SVGD (3.3) for unit time with target density $\pi^{(\ell)}$ incurs a cost c_ℓ. Let each $\pi^{(\ell)} \in C^1(\Theta)$ be continuously differentiable and be such that we may evaluate its score function $\nabla \log \pi^{(\ell)}$. Further, let $\pi^{(\ell)}(\boldsymbol{\theta}) \to \pi(\boldsymbol{\theta})$ for each $\boldsymbol{\theta} \in \Theta$. The pointwise convergence of the densities guarantees convergence of the distributions $\pi^{(\ell)} \to \pi$ in the total variation, and also Hellinger, metrics by Scheffe's lemma. Given a tolerance ϵ and an initial distribution μ_0, the approximation is obtained by first selecting a high-fidelity level $L = L(\epsilon) \in \mathbb{N}$ such that

$$\mathrm{d}_{\mathrm{Hell}}\left(\pi^{(L)},\ \pi\right) \leq \frac{\epsilon}{2}\,. \tag{3.9}$$

This allows us to focus on learning an approximation μ to $\pi^{(L)}$ which is tractable and satisfies

$$\mathrm{d}_{\mathrm{Hell}}\left(\mu,\ \pi^{(L)}\right) \leq \frac{\epsilon}{2}\,.$$

The triangle-inequality for the Hellinger distance will then guarantee that

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

as desired. Note that the Hellinger distance is appealing because it admits a triangle-inequality allowing us to separate the deterministic error due to the fidelity L and the statistical error from the learned approximation μ. Moreover, the Hellinger distance satisfies the following upper bound

$$2\mathrm{d}_{\mathrm{Hell}}\left(\mu_1,\ \mu_2\right)^2 \leq \mathrm{KL}\left(\mu_1 \,||\, \mu_2\right)\,, \tag{3.10}$$

which is convenient for SVGD in particular which minimizes the KL divergence (3.4).

3.2.1 Single-level SVGD

One option to obtain a density μ^{SL} with

$$\mathrm{d}_{\mathrm{Hell}}\left(\mu^{\mathrm{SL}},\ \pi^{(L)}\right) \leq \frac{\epsilon}{2}\,, \tag{3.11}$$

which we refer to as the (continuous-time) single-level SVGD approximation, is to apply SVGD to the high-fidelity target density $\pi^{(L)}$ by solving (3.3) up to time

$$T_{\mathrm{SL}}(\epsilon) = \inf\left\{t \geq 0 : \mathrm{d}_{\mathrm{Hell}}\left(\mu_t,\ \pi^{(L)}\right) \leq \frac{\epsilon}{2}\right\}\,, \tag{3.12}$$

so that

$$\mu^{\mathrm{SL}} = \mu_{T_{\mathrm{SL}}} .$$

Note that the time $T_{\mathrm{SL}} = T_{\mathrm{SL}}(\epsilon)$ will necessarily depend on the initial density μ_0 as well and will increase as $\epsilon \to 0$. The computational cost of single-level SVGD to obtain the approximation μ^{SL} is therefore

$$c_{\mathrm{SL}}(\epsilon) = c_{L(\epsilon)} T_{\mathrm{SL}}(\epsilon) . \tag{3.13}$$

3.2.2 MLSVGD

MLSVGD uses the high-fidelity density $\pi^{(L)}$, with L the same as in (3.9), as well the lower fidelity surrogate densities $\pi^{(1)}, \ldots, \pi^{(L-1)}$ as opposed to single-level SVGD which only uses the high-fidelity density. Because single-level SVGD only integrates with respect to the high-fidelity density $\pi^{(L)}$, which may be computationally expensive, poor initial densities μ_0 may result in a large integration time T_{SL} leading to large a computational cost c_{SL}. MLSVGD aims to circumvent this issue by using the surrogate densities $(\pi^{(\ell)})_{\ell=1}^{L-1}$ to learn increasingly better initial distributions. At the first level $\ell = 1$ we start with the initial density μ_0 and integrate (3.3) with respect to the lowest fidelity surrogate density $\pi^{(1)}$ for time T_1 to obtain an approximation $\mu_{T_1}^{(1)}$ of $\pi^{(1)}$. Then for the next level $\ell = 2$ we start with the new initial density $\mu_{T_1}^{(1)}$ and integrate (3.3) with respect to $\pi^{(2)}$ for time T_2 to obtain $\mu_{T_2^{(2)}}$. In general, for levels $\ell = 2, \ldots, L$ we start with the initial density $\mu_{T_{\ell-1}}^{(\ell-1)}$ and integrate (3.3) with respect to $\pi^{(\ell)}$ for time T_ℓ to obtain the initial density $\mu_{T_\ell}^{(\ell)}$ for the next level. The process terminates at the high-fidelity level L at which point $\mu_{T_L}^{(L)}$ serves as an approximation to $\pi^{(L)}$. We refer to this final approximation as the (continuous-time) MLSVGD approximation

$$\mu^{\mathrm{ML}} = \mu_{T_L}^{(L)} .$$

Figure 3.1 presents a schematic of the procedure for MLSVGD in comparison to the procedure followed by single-level SVGD. Note that the integration times $T_1, \ldots, T_L$ at each level must be chosen to guarantee that the MLSVGD approximation satisfies

$$\mathrm{d}_{\mathrm{Hell}}\left(\mu^{\mathrm{ML}},\ \pi^{(L)}\right) \leq \frac{\epsilon}{2}\,, \tag{3.14}$$

which motivates us to recursively define

$$T_\ell = \inf\left\{t \geq 0\ :\ \mathrm{KL}\left(\mu_t^{(\ell)}\ ||\ \pi^{(\ell)}\right) \leq \frac{\epsilon_\ell^2}{2}\right\}\,, \tag{3.15}$$

for $\ell = 1, \ldots, L$ and where $\epsilon_1 \geq \epsilon_2 \geq \ldots \geq \epsilon_L$ and $\epsilon_L \leq \epsilon$ is a sequence of tolerances and $\mu_t^{(\ell)}$ is the solution of (3.3) at time t with initial density $\mu_{T_{\ell-1}}^{(\ell-1)}$. At the final level $\epsilon_L \leq \epsilon$, which by (3.10) guarantees that the MLSVGD approximation satisfies (3.14). The cost complexity of deriving the MLSVGD approximation μ^{ML} then becomes

$$c_{\mathrm{ML}}(\epsilon) = \sum_{\ell=1}^{L} c_\ell T_\ell\,, \tag{3.16}$$

where both L and $T_1, \ldots, T_L$ will depend on ϵ. Note that by changing the sequence of tolerances $\epsilon_1, \ldots, \epsilon_L$ the integration times (3.15) at each level will necessarily change as well. The motivation behind allowing different tolerances at each level comes from the fact that the objective is to derive an approximation for the high-fidelity density $\pi^{(L)}$ but we are agnostic to whether the intermediate approximations $\mu_{T_\ell}^{(\ell)}$ are accurate. This means that if a surrogate density $\pi^{(\ell)}$ is not close in KL divergence to the high-fidelity density $\pi^{(L)}$ then we do not need to integrate (3.3) for a long time to derive a close approximation $\mu_{T_\ell}^{(\ell)}$ to $\pi^{(\ell)}$ since it will be of limited use for approximating $\pi^{(L)}$. This suggests setting each tolerance ϵ_ℓ to correspond to how close the surrogate density $\pi^{(\ell)}$ is to π in the KL divergence, which we make precise in Proposition 4.

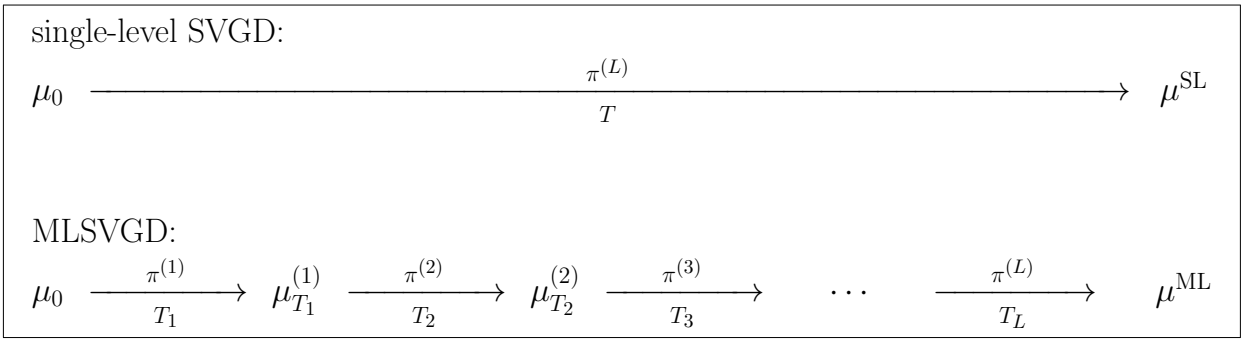


Figure 3.1: An illustration of the procedure for single-level SVGD which only uses the high-fidelity density $\pi^{(L)}$ as opposed to the procedure for MLSVGD which uses the intermediate densities $\pi^{(1)}, \ldots, \pi^{(L-1)}$ sequentially.

3.2.3 Cost-complexity of single-level SVGD and MLSVGD

Single-level SVGD

The cost complexity of single-level SVGD (3.13) depends on the tolerance ϵ, the cost c_L of integrating with respect to the high-fidelity density model $\pi^{(L)}$, and how quickly the SVGD converges to the target distribution. Each of the following three assumptions addresses one of these points and together allow us to derive an upper bound on the cost complexity (3.13) of single-level SVGD in terms ϵ and relevant constants.

Assumption 7. The costs c_ℓ of integrating (3.3) with target density $\pi^{(\ell)}$ for any unit time interval 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 8. There exist constants $\alpha, k_0, k_1 > 0$ independent of ℓ such that

$$\mathrm{KL}\left(\mu_0 \,||\, \pi^{(\ell)}\right) \leq k_0,$$

for all $\ell \in \mathbb{N}$ and

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

where s is the same constant independent of ℓ as in Assumption 7 and μ_0 is the initial distribution.

Assumption 9. There exists a decreasing function $r : [0, \infty) \to [0, 1]$ such that $r(0) = 1$, $\lim_{t\to\infty} r(t) = 0$, and for an initial distribution ν_0

$$\mathrm{KL}\left(\nu_t \,||\, \pi^{(\ell)}\right) \leq r(t)\mathrm{KL}\left(\nu_0 \,||\, \pi^{(\ell)}\right), \qquad \ell \in \mathbb{N},$$

holds, where ν_t is the solution of the nonlinear Fokker-Planck equation (3.3) at time t.

The first assumption 7 is a typical assumption in the multilevel literature (see Section 1.1.3) that allows us to compare the costs of different fidelity models. The second assumption 8 typically goes together with Assumption 7 and captures how quickly the surrogate models converge to the intractable target density π and will allow us to select the high-fidelity level L so that (3.9) is satisfied. The final assumption 9 is independent of the first two assumptions and will guarantee that the single-level SVGD approximation is sufficiently accurate so that (3.11) is satisfied. Note that both Assumptions 8 and 9 use bounds on the KL divergence as opposed to the Hellinger distance directly, which is needed for (3.9) and (3.11). This is advantageous for two reasons. The first is that the Hellinger distance may be upper bounded by the KL divergence so that convergence in the KL divergence implies convergence in the Hellinger distance (3.10). The second is that the KL divergence is much more amenable to the convergence theory of SVGD for the solution of (3.3). In particular, note that for any fixed $\ell \in \mathbb{N}$ the bound in Assumption 9 is guaranteed to hold following from (3.4). Assumption 9 requires that such a function r exists and that this bound holds uniformly for each level $\ell \in \mathbb{N}$. An example of a function r that measures the convergence of SVGD to the target density, which was considered in [Alsup et al., 2021], is the expo-

nential rate $r(t) = \mathrm{e}^{-\lambda t}$ for $\lambda > 0$. It was shown in [Korba et al., 2020] that the bound in Assumption 9 is satisfied with an exponential rate if the distributions $\pi^{(\ell)}$ satisfy a Stein log-Sobolev inequality. Moreover, [Chewi et al., 2020] showed that SVGD converges in the KL divergence with an exponential rate for a specific choice of the kernel K.

Remark 9. There is a strong connection between SVGD and the Langevin algorithm. First note that the Langevin algorithm satisfies a linear Fokker-Planck equation for the density and converges at an exponential rate in the KL divergence when the target measure satisfies a log-Sobolev inequality [Bakry et al., 2014, Theorem 5.2.1]. SVGD on the other hand satisfies a nonlinear Fokker-Planck equation and converges at an exponential rate under a Stein log-Sobolev inequality [Korba et al., 2020]. The exponential rate comes from the gradient flow structure of both the Langevin algorithm and SVGD which gives rise to their corresponding Fokker-Planck equations. We refer to [Jordan et al., 1998] for the connection between gradient dynamics in the space of probability measures and the Fokker-Planck equation. The main difference between the Langevin algorithm and SVGD is that for SVGD the gradient is restricted to be in a RKHS.

Under these assumptions, Proposition 2 derives the cost complexity of obtaining the single-level SVGD approximation.

Proposition 2. *If Assumptions 7,8,9 hold, then the costs of continuous-time SVGD to obtain* μ^{SL} *with*

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

is bounded as

$$\mathrm{c}_{\mathrm{SL}}(\epsilon) \leq c_0 s^{\gamma L} T_{\mathrm{SL}} \leq 2c_0 s^{\gamma} (2k_1)^{\gamma/\alpha} r^{-1} \left(\frac{\epsilon^2}{2\mathrm{KL}\left(\mu_0 \mid\mid \pi^{(L)}\right)}\right) \epsilon^{-2\gamma/\alpha}. \tag{3.17}$$

In the case where r from Assumption 9 is not invertible, which is needed for the cost complexity (3.17), we define

$$r^{-1}(\epsilon) = \inf\left\{t \in [0,\infty) \ : \ r(t) \leq \epsilon\right\},$$

which always exists and is decreasing as $\epsilon \to \infty$, or equivalently increasing as $\epsilon \to 0$.

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

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

so we will bound both of these terms independently by $\epsilon/2$. By inequality (3.10), it is sufficient to bound the KL divergence because

$$\mathrm{d}_{\mathrm{Hell}}\left(\mu^{\mathrm{SL}},\ \pi^{(L)}\right) \leq \sqrt{\frac{\mathrm{KL}\left(\mu^{\mathrm{SL}} \mid\mid \pi^{(L)}\right)}{2}}, \tag{3.18}$$

and similarly for $\mathrm{d}_{\mathrm{Hell}}\left(\pi^{(L)},\ \pi\right)$. By Assumption 8 choose L to be

$$L = \left\lceil \frac{1}{\alpha}\log_s\left(\frac{2k_1}{\epsilon^2}\right)\right\rceil \leq \frac{1}{\alpha}\log_s\left(\frac{2k_1}{\epsilon^2}\right) + 1, \tag{3.19}$$

so that

$$\mathrm{d}_{\mathrm{Hell}}\left(\pi^{(L)},\ \pi\right) \leq \sqrt{\frac{\mathrm{KL}\left(\pi^{(L)} \mid\mid \pi\right)}{2}} \leq \sqrt{\frac{k_1 s^{-\alpha L}}{2}} \leq \frac{\epsilon}{2}. \tag{3.20}$$

The time needed to integrate with SVGD to achieve $\mathrm{d}_{\mathrm{Hell}}\left(\mu^{\mathrm{SL}},\ \pi^{(L)}\right) \leq \epsilon/2$ is

$$T_{\mathrm{SL}} = \inf\left\{t \geq 0 \ : \ \mathrm{d}_{\mathrm{Hell}}\left(\mu_t,\ \pi^{(L)}\right) \leq \frac{\epsilon}{2}\right\}.$$

Again by inequality (3.10),

$$T_{\mathrm{SL}} \leq \inf \left\{ t \geq 0 \ : \ \mathrm{KL}\left(\mu_t \mid\mid \pi^{(L)}\right) \leq \frac{\epsilon^2}{2} \right\}.$$

Now by Assumptions 9, the rate function r is invertible and the time needed to integrate with SVGD to achieve $\mathrm{d}_{\mathrm{Hell}}\left(\mu^{\mathrm{SL}},\ \pi^{(L)}\right) \leq \epsilon/2$ is bounded as

$$T_{\mathrm{SL}} \leq r^{-1}\left(\frac{\epsilon^2}{2\mathrm{KL}\left(\mu_0 \mid\mid \pi^{(L)}\right)}\right). \tag{3.21}$$

With Assumption 7, the total cost to integrate until time T_{SL} at level L is therefore bounded as

$$\mathrm{c}_{\mathrm{SL}}(\epsilon) \leq c_0 s^{\gamma L} T_{\mathrm{SL}} \leq 2c_0 s^{\gamma} (2k_1)^{\gamma/\alpha} r^{-1}\left(\frac{\epsilon^2}{2\mathrm{KL}\left(\mu_0 \mid\mid \pi^{(L)}\right)}\right) \epsilon^{-2\gamma/\alpha}.$$

□

The cost complexity of single-level SVGD (3.17) derived in Proposition 2 directly shows that if the initial distribution μ_0 is far from the high-fidelity distribution $\pi^{(L)}$ in terms of the KL divergence, then a long integration time T_{SL} will be required to converge within the tolerance ϵ resulting in a large computational cost. Moreover, the computational complexity depends on how quickly SVGD converges through the inverse of the rate function r^{-1} from Assumption 9. If SVGD converges slowly, then r will decay slowly as well and hence r^{-1} will be large even for moderate tolerances ϵ.

MLSVGD

Because MLSVGD makes use of the cheaper surrogate densities $\pi^{(1)}, \ldots, \pi^{(L-1)}$ to find a good initial density $\mu_{T_{L-1}}^{(L-1)}$ before integrating (3.3) with respect to the high-fidelity density $\pi^{(L)}$, it can avoid a large integration time T_L at the high-fidelity level and reduce the overall computational cost. MLSVGD iterates over each level, repeatedly performing SVGD. Thus

analogous to (3.12), the integration time T_ℓ required at level ℓ depends on the KL divergence $\mathrm{KL}\left(\mu_{T_{\ell-1}}^{(\ell-1)} \,||\, \pi^{(\ell)}\right)$. Bounding this KL divergence will allow us to bound the times T_ℓ and hence the cost complexity for MLSVGD. To do so, we make the following additional assumption to ensure that the KL divergences between successive densities converges.

Assumption 10. There exists a constant $k_2 > 0$ independent of ℓ such that

$$\mathrm{KL}\left(\pi^{(\ell-1)} \,||\, \pi^{(\ell)}\right) \leq k_2 s^{-\alpha\ell}\,,$$

where $s > 1$ and $\alpha > 0$ are the same constants as in Assumption 8.

Note that Assumption 10 is not immediately implied by Assumption 8 because the KL divergence is not symmetric and does not satisfy the triangle inequality. For example, the surrogate densities $\pi^{(\ell)}$ may all be absolutely continuous with respect to π, but it is possible that for a level ℓ, $\pi^{(\ell-1)}$ is not absolutely continuous with respect to $\pi^{(\ell)}$ resulting in an infinite KL divergence. However, the KL divergence can still be decomposed similar to the triangle inequality, which will be the key result in bounding the KL divergences $\mathrm{KL}\left(\mu_{T_{\ell-1}}^{(\ell-1)} \,||\, \pi^{(\ell)}\right)$. In particular, we have

$$\mathrm{KL}\left(\mu_{T_{\ell-1}}^{(\ell-1)} \,||\, \pi^{(\ell)}\right) = \mathrm{KL}\left(\mu_{T_{\ell-1}}^{(\ell-1)} \,||\, \pi^{(\ell-1)}\right) + \mathrm{KL}\left(\pi^{(\ell-1)} \,||\, \pi^{(\ell)}\right) + R_\ell\,, \tag{3.22}$$

with the remainder R_ℓ given by

$$R_\ell = \int_{\mathbb{R}^d} \left(\mu_{T_{\ell-1}}^{(\ell-1)}(\boldsymbol{\theta}) - \pi^{(\ell-1)}(\boldsymbol{\theta})\right) \log\left(\frac{\pi^{(\ell-1)}(\boldsymbol{\theta})}{\pi^{(\ell)}(\boldsymbol{\theta})}\right) \,\mathrm{d}\boldsymbol{\theta}\,. \tag{3.23}$$

Since $\pi^{(\ell)}(\boldsymbol{\theta}) \to \pi(\boldsymbol{\theta})$ pointwise for $\boldsymbol{\theta} \in \Theta$, $\log \pi^{(\ell-1)}(\boldsymbol{\theta})/\pi^{(\ell)}(\boldsymbol{\theta}) \to 0$ as well and the remainder $R_\ell \to 0$ under mild conditions. The following proposition provides a bound on the cost complexity (3.16) of MLSVGD whenever the remainder term $R_\ell \leq R$ is bounded by a constant R, while Proposition 4 provides a bound on the cost complexity under the slightly

stronger assumption where $R_\ell \to 0$ at the same rate $s^{-\alpha\ell}$ as in Assumptions 8 and 10.

Proposition 3. *If Assumptions 7, 8, 9, and 10 hold, and the remainder in (3.23) $R_\ell \leq R$ is bounded, and the sequence $\epsilon_\ell = \epsilon$ for $\ell = 1, \ldots, L$ in the definition (3.15) of T_ℓ, then continuous-time MLSVGD gives μ^{ML} with $\mathrm{d}_{\mathrm{Hell}}\left(\mu^{\mathrm{ML}},\ \pi\right) \leq \epsilon$ with costs bounded as*

$$c_{\mathrm{ML}}(\epsilon) \leq \frac{c_0 s^{2\gamma}(2k_1)^{\gamma/\alpha}}{s^\gamma - 1} r^{-1} \left(\frac{1}{1 + 2(k_2 + R)\epsilon^{-2}}\right) \epsilon^{-2\gamma/\alpha}\,. \tag{3.24}$$

Proof. As in Equation (3.19) in the proof of Proposition 2 we select the level L as

$$L = \left\lceil \frac{1}{\alpha} \log_s \left(\frac{2k_1}{\epsilon^2}\right) \right\rceil \leq \frac{1}{\alpha} \log_s \left(\frac{2k_1}{\epsilon^2}\right) + 1, \tag{3.25}$$

so that $\mathrm{d}_{\mathrm{Hell}}\left(\pi^{(L)},\ \pi\right) \leq \epsilon/2$. By Assumption 7, the total cost for MLSVGD is bounded by

$$c_{\mathrm{ML}}(\epsilon) \leq \sum_{\ell=1}^{L} c_0 s^{\gamma\ell} T_\ell\,, \tag{3.26}$$

where it remains to bound the integration times T_ℓ at each level. By Assumption 9 and Equation (3.22), we have

$$\begin{aligned} \mathrm{KL}\left(\mu_{T_\ell}^{(\ell)} \,||\, \pi^{(\ell)}\right) &\leq r(T_\ell)\mathrm{KL}\left(\mu_{T_{\ell-1}}^{(\ell-1)} \,||\, \pi^{(\ell)}\right) \\ &= r(T_\ell)\left(\mathrm{KL}\left(\mu_{T_{\ell-1}}^{(\ell-1)} \,||\, \pi^{(\ell-1)}\right) + \mathrm{KL}\left(\pi^{(\ell-1)} \,||\, \pi^{(\ell)}\right) + R_\ell\right)\,, \end{aligned} \tag{3.27}$$

giving a recursive bound on the KL divergence in terms of the KL divergence at the previous level. By the definition (3.15) of the integration times T_ℓ at level ℓ, we know that

$$\mathrm{KL}\left(\mu_{T_\ell}^{(\ell)} \,||\, \pi^{(\ell)}\right) \leq \frac{\epsilon_\ell^2}{2} \tag{3.28}$$

is satisfied for each level $\ell = 1, \ldots, L$. Using (3.28) at level $\ell - 1$ gives

$$\mathrm{KL}\left(\mu_{T_\ell}^{(\ell)} \mid\mid \pi^{(\ell)}\right) \leq r(T_\ell)\left(\frac{\epsilon_{\ell-1}^2}{2} + \mathrm{KL}\left(\pi^{(\ell-1)} \mid\mid \pi^{(\ell)}\right) + R_\ell\right). \tag{3.29}$$

Note that by (3.28) we know that the left-hand-side of (3.29) is guaranteed to be bounded below $\epsilon_\ell^2/2$, but the same is not necessarily true for the right-hand-side which is an upper bound. Instead define T'_ℓ as

$$T'_\ell = \inf\left\{t \geq 0 \;:\; r(t)\left(\frac{\epsilon_{\ell-1}^2}{2} + \mathrm{KL}\left(\pi^{(\ell-1)} \mid\mid \pi^{(\ell)}\right) + R_\ell\right) \leq \frac{\epsilon_\ell^2}{2}\right\}, \tag{3.30}$$

for each level $\ell = 1, \ldots, L$. By (3.29) we know that $T_\ell \leq T'_\ell$. Solving directly gives

$$T'_\ell = r^{-1}\left(\frac{\epsilon_\ell^2}{\epsilon_{\ell-1}^2 + 2\mathrm{KL}\left(\pi^{(\ell-1)} \mid\mid \pi^{(\ell)}\right) + 2R_\ell}\right). \tag{3.31}$$

We now use the facts that r^{-1} is decreasing, $\epsilon_\ell = \epsilon$ for $\ell = 1, \ldots, L$, $R_\ell \leq R$, and $\mathrm{KL}\left(\pi^{(\ell-1)} \mid\mid \pi^{(\ell)}\right) \leq k_2 s^{-\alpha\ell} \leq k_2$ to obtain

$$T_\ell \leq r^{-1}\left(\frac{\epsilon^2}{\epsilon^2 + 2k_2 + 2R}\right). \tag{3.32}$$

Therefore, the total cost can be bounded by

$$c_{\mathrm{ML}}(\epsilon) \leq \sum_{\ell=1}^{L} c_0 s^{\gamma\ell} r^{-1}\left(\frac{\epsilon^2}{\epsilon^2 + 2k_2 + 2R}\right). \tag{3.33}$$

Since the terms in this sum correspond to a geometric series, we can compute the sum in (3.33) exactly

$$c_{\mathrm{ML}}(\epsilon) \leq c_0 s^\gamma r^{-1}\left(\frac{\epsilon^2}{\epsilon^2 + 2k_2 + 2R}\right)\frac{s^{\gamma L} - 1}{s^\gamma - 1} \leq c_0 s^\gamma r^{-1}\left(\frac{\epsilon^2}{\epsilon^2 + 2k_2 + 2R}\right)\frac{s^{\gamma L}}{s^\gamma - 1}, \tag{3.34}$$

where we have added 1 in the numerator on the right-hand-side only for convenience. Plugging in the upper bound (3.19) on the level L and simplifying terms gives the final upper bound on the cost complexity of the continuous-time MLSVGD approximation μ^{ML}

$$c_{\mathrm{ML}}(\epsilon) \leq \frac{c_0 s^{2\gamma}(2k_1)^{\gamma/\alpha}}{s^{\gamma}-1} r^{-1}\left(\frac{1}{1+2(k_2+R)\epsilon^{-2}}\right)\epsilon^{-2\gamma/\alpha}\,. \tag{3.35}$$

□

First note that the proof of Proposition 3 is presented slightly differently from the proof in [Alsup et al., 2021] in order to elucidate the dependence of the cost complexity on Assumption 10 and the assumption that $R_\ell \leq R$. When $\epsilon \to 0$ the term

$$r^{-1}\left(\frac{1}{1+2(k_2+R)\epsilon^{-2}}\right)$$

behaves asymptotically as

$$r^{-1}\left(\frac{\epsilon^2}{2(k_2+R)}\right),$$

which is analogous to the term

$$r^{-1}\left(\frac{\epsilon^2}{2\mathrm{KL}\left(\mu_0 \,||\, \pi^{(L)}\right)}\right)$$

from the upper bound on the cost complexity (3.17) of single-level SVGD arising from bounding the integration time T_{SL}. This means that asymptotically the cost complexities for MLSVGD and single-level SVGD may grow at the same rate with $k_2 + R$ analogous to $\mathrm{KL}\left(\mu_0 \,||\, \pi\right)$. However, we used three crude approximations throughout the proof of Proposition 3. First is that we only used Assumption 10 to obtain a bound $\mathrm{KL}\left(\pi^{(\ell-1)} \,||\, \pi^{(\ell)}\right)$, which is actually much weaker than what Assumption 10 required. Second is that the remainder term R_ℓ is only bounded, when in fact it will typically converge at a rate $s^{-\alpha\ell}$ as

well. We show that this happens in the Bayesian inverse problem setting presented later in Section 3.4. Third and finally, we set the sequence of tolerances $\epsilon_\ell = \epsilon$ to be fixed. However, this defeats the motivation behind the choice of tolerances ϵ discussed earlier. A better choice is to select ϵ_ℓ to be decreasing and correspond to the KL divergences at each level. Proposition 4 that follows shows that MLSVGD may have a much more favorable cost complexity compared to single-level SVGD when these mild assumptions are met.

Proposition 4. *If Assumptions 7, 8, 10, and 9 hold and $R_\ell \leq k_3 s^{-\alpha\ell}$, then by setting $\epsilon_\ell = \sqrt{2k_1}s^{-\alpha\ell/2}$ for $\ell = 1, \ldots, L$, the costs of continuous-time MLSVGD to have $\mathrm{d}_{\mathrm{Hell}}\left(\mu^{\mathrm{ML}},\ \pi\right) \leq \epsilon$ can be bounded as*

$$c_{\mathrm{ML}}(\epsilon) \leq \frac{c_0 s^{2\gamma}(2k_1)^{\gamma/\alpha}}{s^\gamma - 1} r^{-1}\left(\frac{1}{s^\alpha + (k_2+k_3)/k_1}\right)\epsilon^{-2\gamma/\alpha}\,. \tag{3.36}$$

Proof. First note that by setting $\epsilon_\ell = \sqrt{2k_1}s^{-\alpha\ell/2}$ for $\ell = 1, \ldots, L$ in (3.15) we have that

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

by the choice of the high-fidelity level L (3.19). Proceeding as in the proof of Proposition 3 we use Assumption 10 and the new assumption that $R_\ell \leq k_3 s^{-\alpha\ell}$ to replace the bound (3.32) on T_ℓ with

$$T_\ell = r^{-1}\left(\frac{2k_1 s^{-\alpha\ell}}{2k_1 s^\alpha s^{-\alpha\ell} + 2k_2 s^{-\alpha\ell} + 2k_3 s^{-\alpha\ell}}\right)\,,$$

where we have directly plugged in ϵ_ℓ and $\epsilon_{\ell-1}$. Simplifying gives the bound

$$T_\ell \leq r^{-1}\left(\frac{1}{s^\alpha + (k_2+k_3)/k_1}\right)\,, \tag{3.37}$$

which is independent of the tolerance ϵ. The total cost can now be bounded by

$$c_{\mathrm{ML}}(\epsilon) \leq \sum_{\ell=1}^{L} c_0 s^{\gamma\ell} r^{-1} \left(\frac{1}{s^\alpha + (k_2 + k_3)/k_1} \right), \tag{3.38}$$

which we may again compute explicitly

$$\begin{aligned} c_{\mathrm{ML}}(\epsilon) &\leq c_0 s^\gamma r^{-1} \left(\frac{1}{s^\alpha + (k_2 + k_3)/k_1} \right) \frac{s^{\gamma L} - 1}{s^\gamma - 1} \\ &\leq c_0 s^\gamma r^{-1} \left(\frac{1}{s^\alpha + (k_2 + k_3)/k_1} \right) \frac{s^{\gamma L}}{s^\gamma - 1}, \end{aligned} \tag{3.39}$$

and we have again added 1 in the numerator of the last term for convenience. Plugging in the upper bound (3.19) on the level L and simplifying terms gives the final upper bound on the improved cost complexity of the continuous-time MLSVGD approximation μ^{ML}

$$c_{\mathrm{ML}}(\epsilon) \leq \frac{c_0 s^{2\gamma} (2k_1)^{\gamma/\alpha}}{s^\gamma - 1} r^{-1} \left(\frac{1}{s^\alpha + (k_2 + k_3)/k_1} \right) \epsilon^{-2\gamma/\alpha}. \tag{3.40}$$

□

The improved MLSVGD cost complexity bound derived in Proposition 4 when the surrogate densities satisfy Assumption 10 and the remainder also converges at the same rate has two notable advantages over the single-level SVGD cost complexity (3.13) or the cost complexity of MLSVGD without these assumptions (3.24). The first notable difference is that the bounds on the integration times T_ℓ (3.37) for MLSVGD in Proposition 4 are independent of the tolerance ϵ. However, we note that as $\epsilon \to 0$ the high-fidelity level L will increase as well and so although the times T_ℓ remain bounded the number of levels that MLSVGD iterates over will increase. A consequence of bounded integration times, is that the cost complexity (3.40) scales only as $\epsilon^{-2\gamma/\alpha}$ as opposed to single-level SVGD which has an additional $r^{-1}(\epsilon^2/2\mathrm{KL}\left(\mu_0 \mid\mid \pi^{(L)}\right))$ factor. If SVGD converges slowly, then this additional

factor may provide a large contribution to the cost complexity of single-level SVGD. In this setting MLSVGD may have longer integration times at the cheaper levels and shorter integration times at the more expensive higher levels. On the other hand, if SVGD converges quickly, then there is little speedup that can be achieved since the integration times will be short regardless. The second notable advantage of MLSVGD over single-level SVGD is that the bound on the cost complexity (3.36) for MLSVGD is independent of the initial KL divergence $\mathrm{KL}\left(\mu_0 \,||\, \pi^{(L)}\right)$, but rather depends on the constants k_2 from Assumption 10 and k_3 from (3.23). The constant k_2 controls the KL divergence between two consecutive levels and the constant k_3 controls how close the SVGD approximation $\mu_{T_{\ell-1}}^{(\ell-1)}$ is to $\pi^{(\ell-1)}$. If k_2 and k_3 are both small, then KL divergence between consecutive levels is small as well and the approximations $\mu_{T_\ell}^{(\ell-1)}$ at each level ℓ serve as good initial densities for the following level, resulting in reduced integration times and costs.

3.3 Discrete-time single-level SVGD and MLSVGD

The analysis presented in Section 3.2 as well as that presented in [Alsup et al., 2021] only applies to continuous-time MLSVGD and continuous-time single-level SVGD obtained by solving the nonlinear Fokker-Planck equation (3.3). However, drawing samples from either μ^{SL} or μ^{ML} in this setting requires integrating the characteristic flow (3.2) exactly, which may not be practical. In this section, we show that the analysis from the continuous-time setting may be carried over with minimal adjustments to the discrete-time setting where the particles are updated iteratively according to the discretization (3.5) of the characteristic flow (3.2).

3.3.1 Discrete-time notation and modifications

The discrete-time single-level SVGD approximation is defined similarly to the continuous-time single-level SVGD approximation. Given the tolerance ϵ we select the high-fidelity level L the same as in (3.19) so that

$$\mathrm{d}_{\mathrm{Hell}}\left(\pi^{(L)},\ \pi\right) \leq \frac{\epsilon}{2},$$

which is independent of the SVGD approximation, whether discrete or continuous in time. Next we define

$$T_{\mathrm{SL}} = \inf\left\{\tau \geq 0 : \mathrm{d}_{\mathrm{Hell}}\left(\mu_\tau,\ \pi^{(L)}\right) \leq \frac{\epsilon}{2}\right\}, \tag{3.41}$$

where μ_τ is given by the updates (3.6). The only differences from the continuous-time setting is that the infimum is taken over integers τ and that the density μ_τ is obtained from (3.6) as opposed to solving (3.3). Due to the near-identical definitions we overload our notation and also write $\mu^{\mathrm{SL}} = \mu_{T_{\mathrm{SL}}}$ to denote the discrete-time single-level SVGD approximation, where it will be clear from context whether μ^{SL} refers to the continuous-time or discrete-time version. The discrete-time MLSVGD definition is analgous to the discrete-time single-level SVGD approximation with the only difference being the definition of the integration times T_ℓ at each level

$$T_\ell = \inf\left\{\tau \geq 0\ :\ \mathrm{KL}\left(\mu_\tau^{(\ell)}\ ||\ \pi^{(\ell)}\right) \leq \frac{\epsilon_\ell^2}{2}\right\}, \tag{3.42}$$

where $\mu_\tau^{(\ell)}$ is the density given at iteration τ of (3.6) when starting from distribution $\mu_{T_{\ell-1}}^{(\ell-1)}$ and updating with respect to the surrogate density $\pi^{(\ell)}$.

3.3.2 Cost complexity for discrete-time versions

Using the overloaded notation in the previous section we can trivially extend the results of Propositions 2, 3, and 4 to the discrete time setting. To do so we make two minor modifications to the assumptions presented earlier to be compatible with the discrete time

setting. The first change is to Assumption 7, which now concerns the cost of evaluating the score functions surrogate densities as opposed to integrating with respect to them. The second change is regarding Assumption 9, which is again primarily notational to instead consider functions r defined over the non-negative integers $\mathbb{N}_0$.

Assumption 11. The costs c_ℓ of evaluating the score functions $\nabla \log \pi^{(\ell)}(\boldsymbol{\theta})$ at any point $\boldsymbol{\theta} \in \Theta$, are bounded as

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

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

Assumption 12. There exists a strictly decreasing function $r : \mathbb{N}_0 \to [0, 1]$ such that $r(0) = 1$, $\lim_{\tau \to \infty} r(\tau) = 0$, and for any initial distribution ν_0

$$\mathrm{KL}\left(\nu_\tau \,||\, \pi^{(\ell)}\right) \leq r(\tau) \mathrm{KL}\left(\nu_0 \,||\, \pi^{(\ell)}\right), \qquad \ell \in \mathbb{N},$$

holds, where ν_τ evolves according to the discrete-time SVGD update (3.5).

Just as in the continuous-time setting where (3.4) guarantees that the KL divergence of the solution to (3.3) decreases over time, the discrete analogue (3.7) guarantees that the density updates μ_τ of (3.6) also decrease the KL divergence. Since the rate function r in Assumption 12 is now a function on the non-negative integers it may no longer have an inverse defined on $[0, 1]$. Instead we define

$$r^{-1}(\epsilon) = \inf \{\tau \geq 0 : r(\tau) \leq \epsilon\},$$

which is again notationally the same as in the continuous-time case. We now can state the cost complexity bounds for the discrete-time versions of single-level SVGD and MLSVGD. With the overloaded notation presented in this section, the proofs of each of the following three propositions are identical to their continuous-time counterparts. In particular, we see

that the same cost complexities whether the discrete or continuous time setting is used.

Proposition 5. *If Assumptions 11,8,12 hold, then the cost of discrete-time single-level SVGD to obtain* μ^{SL} *with*

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

is bounded as

$$\mathrm{c}_{\mathrm{SL}}(\epsilon) \leq c_0 s^{\gamma L} T_{\mathrm{SL}} \leq 2c_0 s^{\gamma} (2k_1)^{\gamma/\alpha} r^{-1} \left(\frac{\epsilon^2}{2\mathrm{KL}\left(\mu_0 \mid\mid \pi^{(L)}\right)}\right) \epsilon^{-2\gamma/\alpha}. \tag{3.43}$$

Proposition 6. *If Assumptions 11, 8, 12, and 10 hold, and the remainder in (3.23)* $R_\ell \leq R$ *is bounded, and the sequence* $\epsilon_\ell = \epsilon$ *for* $\ell = 1, \ldots, L$ *in the definition (3.42) of* T_ℓ*, then discrete-time MLSVGD gives* μ^{ML} *with* $\mathrm{d}_{\mathrm{Hell}}\left(\mu^{\mathrm{ML}},\ \pi\right) \leq \epsilon$ *with costs bounded as*

$$c_{\mathrm{ML}}(\epsilon) \leq \frac{c_0 s^{2\gamma} (2k_1)^{\gamma/\alpha}}{s^{\gamma} - 1} r^{-1} \left(\frac{1}{1 + 2(k_2 + R)\epsilon^{-2}}\right) \epsilon^{-2\gamma/\alpha}. \tag{3.44}$$

Proposition 7. *If Assumptions 11, 8, 10, and 12 hold and* $R_\ell \leq k_3 s^{-\alpha\ell}$*, then by setting* $\epsilon_\ell = \sqrt{2k_1} s^{-\alpha\ell/2}$ *for* $\ell = 1, \ldots, L$*, the costs of discrete-time MLSVGD to have* $\mathrm{d}_{\mathrm{Hell}}\left(\mu^{\mathrm{ML}},\ \pi\right) \leq \epsilon$ *can be bounded as*

$$c_{\mathrm{ML}}(\epsilon) \leq \frac{c_0 s^{2\gamma} (2k_1)^{\gamma/\alpha}}{s^{\gamma} - 1} r^{-1} \left(\frac{1}{s^{\alpha} + (k_2 + k_3)/k_1}\right) \epsilon^{-2\gamma/\alpha}. \tag{3.45}$$

Because the cost-complexities and definitions of the single-level SVGD and MLSVGD are the same in both discrete-time and continuous-time, from now on we simply write "single-

level SVGD" (or just SVGD) and "MLSVGD" where the setting is clear from context or does not matter.

3.4 MLSVGD for Bayesian inverse problems

Bayesian inverse problems arising from scientific and engineering applications often depend on the solution of an underlying PDE that describes some physical process. In this case, the forward model G and hence the corresponding posterior density

$$\pi(\boldsymbol{\theta}) = \frac{1}{Z} \exp\left(-\frac{1}{2} \|\mathbf{y} - G(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right) \pi_0(\boldsymbol{\theta}) \,, \tag{3.46}$$

may be intractable to evaluate exactly, c.f. Section 1.3. Furthermore, by discretizing the underlying PDE to approximate the solution operator, we can obtain a hierarchy of surrogate models $(G^{(\ell)})_{\ell \geq 1}$ to obtain the sequence of posterior surrogate densities $(\pi^{(\ell)})_{\ell \geq 1}$. Single-level SVGD and especially MLSVGD are attractive methods that provide theoretical guarantees on the error with respect to the true posterior π while only having to evaluate the surrogate densities $\pi^{(1)}, \ldots, \pi^{(L)}$. The next two assumptions are specific to the Bayesian inverse problem setting and will allow us to recover the cost complexities derived in the previous sections.

Assumption 13 (Model error). There is a function $\psi : \mathbb{N} \to (0, \infty)$, with $\psi(\ell) \to 0$ as $\ell \to \infty$, such that

$$\left\|G(\boldsymbol{\theta}) - G^{(\ell)}(\boldsymbol{\theta})\right\|_{L^2(\pi_0)} \leq \psi(\ell) \,, \tag{3.47}$$

where the $\|\cdot\|_{L^2(\pi_0)}$ is the L^2 norm over π_0.

Assumption 14. There exists a constant $b_3 > 0$ independent of ℓ such that

$$\mu^{(\ell)}_{T_\ell}(\boldsymbol{\theta}) \leq b_3 \pi_0(\boldsymbol{\theta}) \tag{3.48}$$

for all $\ell \geq 1$.

Using Assumption 13, the next two lemmas will translate the bound on the L^2 error of the surrogate models to a bound on the KL divergence of the surrogate posterior density needed for Assumptions 8 and 10. Thus, in the Bayesian inverse problem setting, one may effectively replace Assumptions 8 and 10 with Assumption 13, which may be easier to verify from standard approximation results in numerical analysis. We note the proof of the following lemma closely mirrors the proofs of Lemmas 4.2 and 4.3 in [Marzouk and Xiu, 2009], but is slightly more general.

Lemma 3. *If Assumption 13 holds, there exists a constant $C > 0$ such that for all $1 \leq \ell_1, \ell_2 \leq \infty$ sufficiently large*

$$\mathrm{KL}(\pi^{(\ell_1)} \mid\mid \pi^{(\ell_2)}) \leq C \|G^{(\ell_1)} - G^{(\ell_2)}\|_{L^2(\pi_0)}\,. \tag{3.49}$$

Note that for $\ell = \infty$ we set $G^{(\ell)} = G$.

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 $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ and symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ we have

$$\begin{aligned}
\|\mathbf{u}-\mathbf{w}\|_{\mathbf{A}}^2 - \|\mathbf{v}-\mathbf{w}\|_{\mathbf{A}}^2 &= \|(\mathbf{u}-\mathbf{v})+(\mathbf{v}-\mathbf{w})\|_{\mathbf{A}}^2 - \|\mathbf{v}-\mathbf{w}\|_{\mathbf{A}}^2 \\
&= \langle(\mathbf{u}-\mathbf{v}),\ \mathbf{A}(\mathbf{u}-\mathbf{v})\rangle + 2\langle(\mathbf{u}-\mathbf{v}),\ \mathbf{A}(\mathbf{v}-\mathbf{w})\rangle \\
&= \langle(\mathbf{u}-\mathbf{v}),\ \mathbf{A}(\mathbf{u}+\mathbf{v}-2\mathbf{w})\rangle \\
&\leq \|\mathbf{u}-\mathbf{v}\| \cdot \|\mathbf{A}(\mathbf{u}+\mathbf{v}-2\mathbf{w})\|\,,
\end{aligned} \tag{3.50}$$

with the last line following from the Cauchy-Schwarz inequality. Applying this bound with

$\mathbf{u} = G_1(\boldsymbol{\theta})$, $\mathbf{v} = G_2(\boldsymbol{\theta})$, $\mathbf{w} = \mathbf{y}$, and $\mathbf{A} = \boldsymbol{\Gamma}^{-1}$ gives

$$\begin{aligned}
&\int_\Theta \left| \|\mathbf{y} - G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G_2(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \pi_0(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\
&\leq \int_\Theta \|G_1(\boldsymbol{\theta}) - G_2(\boldsymbol{\theta})\| \cdot \|\boldsymbol{\Gamma}^{-1}(2\mathbf{y} - G_1(\boldsymbol{\theta}) - G_2(\boldsymbol{\theta}))\| \pi_0(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\
&\leq \|G_1 - G_2\|_{L^2(\pi_0)} \cdot \|\boldsymbol{\Gamma}^{-1}(2\mathbf{y} - G_1 - G_2)\|_{L^2(\pi_0)} \,,
\end{aligned} \tag{3.51}$$

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 (3.51)

$$\begin{aligned}
\mathrm{KL}(\pi_1 \,||\, \pi_2) &= \int_\Theta \pi_1(\boldsymbol{\theta}) \log\left(\frac{\pi_1(\boldsymbol{\theta})}{\pi_2(\boldsymbol{\theta})}\right) \, \mathrm{d}\boldsymbol{\theta} \\
&= \int_\Theta \pi_1(\boldsymbol{\theta}) \log\left(\frac{Z_2 \exp\left(-\frac{1}{2}\|\mathbf{y} - G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}{Z_1 \exp\left(-\frac{1}{2}\|\mathbf{y} - G_2(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}\right) \, \mathrm{d}\boldsymbol{\theta} \\
&= \log\left(\frac{Z_2}{Z_1}\right) + \int_\Theta \pi_1(\boldsymbol{\theta}) \log\left(\frac{\exp\left(-\frac{1}{2}\|\mathbf{y} - G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}{\exp\left(-\frac{1}{2}\|\mathbf{y} - G_2(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}\right) \, \mathrm{d}\boldsymbol{\theta} \\
&\leq \log\left(\frac{Z_2}{Z_1}\right) + \frac{1}{2Z_1} \int_\Theta \left| \|\mathbf{y} - G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G_2(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \pi_0(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\
&\leq \left| \log\left(\frac{Z_2}{Z_1}\right) \right| + \frac{1}{2Z_1} \|G_1 - G_2\|_{L^2(\pi_0)} \cdot \|\boldsymbol{\Gamma}^{-1}(2\mathbf{y} - G_1 - G_2)\|_{L^2(\pi_0)} \,,
\end{aligned} \tag{3.52}$$

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

$$\exp\left(-\frac{1}{2}\|\mathbf{y} - G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right) \leq 1 \,. \tag{3.53}$$

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 (3.51)

$$\begin{aligned}
|Z_1 - Z_2| &= \left| \int_\Theta \left\{ \exp\left(-\frac{1}{2} \|\mathbf{y} - G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right) - \exp\left(-\frac{1}{2} \|\mathbf{y} - G_2(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right) \right\} \pi_0(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \right| \\
&\leq \int_\Theta \left| \exp\left(-\frac{1}{2} \|\mathbf{y} - G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right) - \exp\left(-\frac{1}{2} \|\mathbf{y} - G_2(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right) \right| \pi_0(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\
&\leq \frac{1}{2} \int_\Theta \left| \|\mathbf{y} - G_1(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G_2(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \pi_0(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\
&\leq \frac{1}{2} \|G_1 - G_2\|_{L^2(\pi_0)} \cdot \|\boldsymbol{\Gamma}^{-1}(2\mathbf{y} - G_1 - G_2)\|_{L^2(\pi_0)} \, .
\end{aligned} \tag{3.54}$$

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

$$\begin{aligned}
\|\boldsymbol{\Gamma}^{-1}(2\mathbf{y} - G_1 - G_2)\|_{L^2(\pi_0)} &\leq 2\|\boldsymbol{\Gamma}^{-1}\mathbf{y}\|_{L^2(\pi_0)} + \|\boldsymbol{\Gamma}^{-1}(G_1 + G_2)\|_{L^2(\pi_0)} \\
&\leq 2\|\boldsymbol{\Gamma}^{-1}\mathbf{y}\|_{L^2(\pi_0)} + 2\|\boldsymbol{\Gamma}^{-1}G\|_{L^2(\pi_0)} + \|\boldsymbol{\Gamma}^{-1}(G_1 + G_2 - 2G)\|_{L^2(\pi_0)} \\
&\leq 2\|\boldsymbol{\Gamma}^{-1}\mathbf{y}\|_{L^2(\pi_0)} + 2\|\boldsymbol{\Gamma}^{-1}G\|_{L^2(\pi_0)} \\
&\quad + \frac{1}{\gamma_{\min}} \|G_1 - G\|_{L^2(\pi_0)} + \frac{1}{\gamma_{\min}} \|G_2 - G\|_{L^2(\pi_0)} \, .
\end{aligned} \tag{3.55}$$

Since $\|G^{(\ell)} - G\|_{L^2(\pi_0)} \to 0$ by Assumption 13, 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

$$\|\boldsymbol{\Gamma}^{-1}(2\mathbf{y} - G_1 - G_2)\|_{L^2(\pi_0)} \leq b_1. \tag{3.56}$$

Combining Equations (3.54) and (3.56) yields

$$|Z_1 - Z_2| \leq \frac{b_1}{2} \|G_1 - G_2\|_{L^2(\pi_0)} \, . \tag{3.57}$$

The ratio of the normalizing constants can be written

$$\left|\frac{Z_2}{Z_1} - 1\right| = \frac{1}{Z_1}\left|Z_1 - Z_2\right| , \tag{3.58}$$

so the logarithm can be bounded as

$$\left|\log\left(\frac{Z_2}{Z_1}\right)\right| \leq \max\left\{\left|\log\left(1 - \frac{|Z_2 - Z_1|}{Z_1}\right)\right|,\ \log\left(1 + \frac{|Z_2 - Z_1|}{Z_1}\right)\right\} \tag{3.59}$$

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

$$\left|\log\left(\frac{Z_2}{Z_1}\right)\right| \leq \max\left\{\frac{\frac{|Z_2 - Z_1|}{Z_1}}{1 - \frac{|Z_2 - Z_1|}{Z_1}},\ \frac{|Z_2 - Z_1|}{Z_1}\right\} \leq \frac{|Z_1 - Z_2|}{Z_1 - |Z_1 - Z_2|} . \tag{3.60}$$

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

$$Z_1^{-1} \leq \sup_{\ell \geq 1} Z_\ell^{-1} \leq b_2 . \tag{3.61}$$

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

$$\left|\log\left(\frac{Z_2}{Z_1}\right)\right| \leq \frac{|Z_1 - Z_2|}{b_2^{-1} - |Z_1 - Z_2|} \leq 2b_2|Z_1 - Z_2| . \tag{3.62}$$

Combining Equations (3.52), (3.56), (3.57), (3.61), and (3.62) gives

$$\mathrm{KL}(\pi_1 \mid\mid \pi_2) \leq \frac{3}{2} b_1 b_2 \|G_1 - G_2\|_{L^2(\pi_0)} . \tag{3.63}$$

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

Lemma 4. *If Assumption 13 holds with $\psi(\ell) = b_0 s^{-\alpha\ell}$, then Assumptions 8, 10 also hold with the same rate α.*

Proof. Let $\ell_1 = \ell$ and $\ell_2 = \infty$, so that by Lemma 3 we immediately have

$$\mathrm{KL}\left(\pi^{(\ell)} \mid\mid \pi\right) \leq C \left\|G^{(\ell)} - G\right\|_{L^2(\pi_0)} \leq C\psi(\ell) = Cb_0 s^{-\alpha\ell}, \tag{3.64}$$

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

$$\mathrm{KL}\left(\pi^{(\ell-1)} \mid\mid \pi^{(\ell)}\right) \leq C\|G_{\ell-1} - G^{(\ell)}\|_{L^2(\pi_0)} \leq C\left(\|G_{\ell-1} - G\|_{L^2(\pi_0)} + \|G_\ell - G\|_{L^2(\pi_0)}\right). \tag{3.65}$$

Thus,

$$\mathrm{KL}\left(\pi^{(\ell-1)} \mid\mid \pi^{(\ell)}\right) \leq C\left(1 + \frac{\psi(\ell-1)}{\psi(\ell)}\right)\psi(\ell) \leq Cb_0\left(1 + s^\alpha\right)s^{-\alpha\ell}, \tag{3.66}$$

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

Note that Assumption 14 has not yet been used and is only needed in the following theorem to help bound the remainder terms R_ℓ by ensuring that the SVGD densities obtained by solving (3.3) (or equivalently (3.6) in the discrete-time setting) are absolutely continuous with respect to the prior. The next theorem shows that in the Bayesian inverse problem setting, under these new assumptions on the surrogate models and prior, the cost complexities obtained in Sections 3.2 and 3.3 still hold where $(\pi^{(\ell)})_{\ell \geq 1}$ is now the sequence of posterior distributions. Although the next theorem gives an upper bound on the cost complexity for continuous-time MLSVGD, the corresponding result for discrete-time MLSVGD will be the same by replacing Assumption 7 with Assumption 11 and Assumption 9 with Assumption 12. Additionally, we may derive a similar upper bound on the cost complexity for single-level SVGD in the Bayesian inverse problem setting through the same steps.

Theorem 4. *If Assumptions 7, 9 (or equivalently Assumption 11 and 12 in the discrete-time setting), and 14 hold and Assumption 13 holds with $\psi(\ell) = b_0 s^{-\alpha\ell}$, then Assumptions 8*

and 10 hold and thus the cost complexity to find μ^{ML} *with* $\mathrm{d}_{\mathrm{Hell}}\left(\mu^{\mathrm{ML}},\ \pi\right) \leq \epsilon$ *is given by*

$$c_{\mathrm{ML}}(\epsilon) \leq \frac{c_0 s^{2\gamma}(2Cb_0)^{\gamma/\alpha}}{s^{\gamma}-1} r^{-1} \left(\frac{1}{(1+s^{\alpha})\left(2+\frac{b_1 b_2 + b_3}{2C}\right) - 1} \right) \epsilon^{-2\gamma/\alpha}\,, \tag{3.67}$$

where the constants b_1, b_2 *are independent of* ϵ *and given in the proof of Lemma 3.*

Proof. By Lemma 4 we know that Assumptions 8 and 10 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 4.

$$\begin{aligned} R_\ell &= \int_{\Theta} \left(\mu_{T_{\ell-1}}^{(\ell-1)}(\boldsymbol{\theta}) - \pi^{(\ell-1)}(\boldsymbol{\theta})\right) \log\left(\frac{\pi^{(\ell-1)}(\boldsymbol{\theta})}{\pi^{(\ell)}(\boldsymbol{\theta})}\right)\, \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\Theta} \left(\mu_{T_{\ell-1}}^{(\ell-1)}(\boldsymbol{\theta}) - \pi^{(\ell-1)}(\boldsymbol{\theta})\right) \log\left(\frac{Z_\ell \exp\left(-\frac{1}{2}\|\mathbf{y} - G^{(\ell-1)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}{Z_{\ell-1}\exp\left(-\frac{1}{2}\|\mathbf{y} - G^{(\ell)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}\right)\, \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\Theta} \left(\mu_{T_{\ell-1}}^{(\ell-1)}(\boldsymbol{\theta}) - \pi^{(\ell-1)}(\boldsymbol{\theta})\right) \log\left(\frac{\exp\left(-\frac{1}{2}\|\mathbf{y} - G^{(\ell-1)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}{\exp\left(-\frac{1}{2}\|\mathbf{y} - G^{(\ell)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}}\right)}\right)\, \mathrm{d}\boldsymbol{\theta}, \end{aligned} \tag{3.68}$$

where the last line follows from the fact that

$$\int_{\Theta} \left(\mu_{T_{\ell-1}}^{(\ell-1)}(\boldsymbol{\theta}) - \pi^{(\ell-1)}(\boldsymbol{\theta})\right) \log\left(\frac{Z_\ell}{Z_{\ell-1}}\right)\, \mathrm{d}\boldsymbol{\theta} = 0 \tag{3.69}$$

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

$$\begin{aligned} R_\ell \leq \frac{1}{2}\int_{\Theta} \left| \|\mathbf{y} - G^{(\ell)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G^{(\ell-1)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \mu_{T_{\ell-1}}^{(\ell-1)}(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\ + \frac{1}{2}\int_{\Theta} \left| \|\mathbf{y} - G^{(\ell)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G^{(\ell-1)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \pi^{(\ell-1)}(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}\,. \end{aligned} \tag{3.70}$$

We have that

$$\pi^{(\ell-1)}(\boldsymbol{\theta}) \leq \frac{1}{Z_{\ell-1}}\pi_0(\boldsymbol{\theta})\,, \tag{3.71}$$

so that when combined with Assumption 14

$$
\begin{aligned}
R_\ell &\leq \frac{1}{2}\int_\Theta \left| \|\mathbf{y} - G^{(\ell)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G^{(\ell-1)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \mu^{(\ell-1)}_{T_{\ell-1}}(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\
&\quad + \frac{1}{2}\int_\Theta \left| \|\mathbf{y} - G^{(\ell)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G^{(\ell-1)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \pi^{(\ell-1)}(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\
&\leq \frac{b_3}{2}\int_\Theta \left| \|\mathbf{y} - G^{(\ell)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G^{(\ell-1)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \pi_0(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\
&\quad + \frac{1}{2Z_{\ell-1}}\int_\Theta \left| \|\mathbf{y} - G^{(\ell)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} - \|\mathbf{y} - G^{(\ell-1)}(\boldsymbol{\theta})\|^2_{\boldsymbol{\Gamma}^{-1}} \right| \pi_0(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\
&\leq \left(\frac{b_3}{2} + \frac{b_1 b_2}{2}\right) \|G^{(\ell)} - G^{(\ell-1)}\|_{L^2(\pi_0)},
\end{aligned}
\tag{3.72}
$$

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 4 gives the result. □

3.5 A practical MLSVGD algorithm with adaptive stopping criterion

In practice, one wishes to draw samples $\boldsymbol{\theta} \sim \mu^{\mathrm{ML}}$ from the MLSVGD (or SVGD) approximation to perform inference of the intractable distribution π. For this purpose, discrete-time MLSVGD and SVGD are more amenable because they avoid integrating the characteristic flow (3.2) but rather update the particle with a sequence of transport maps (3.5). However, the updates (3.5) require exact computation of the gradient $\nabla J_\mu(\mathbf{0})$, which again may not be feasible. Practical implementations of SVGD, and hence MLSVGD as well, update an ensemble of particles $\{\boldsymbol{\theta}_\tau^{[i]}\}_{i=1}^N$ simultaneously to approximate the gradient $\mathbf{g}_\tau = \nabla J_{\mu_\tau}(\mathbf{0})$ by replacing the expectation in (3.1) with a sample average over the ensemble

$$
\hat{\mathbf{g}}_\tau^{(\ell)}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^N K(\boldsymbol{\theta}_\tau^{[i]}, \boldsymbol{\theta}) \nabla \log \pi^{(\ell)}(\boldsymbol{\theta}_\tau^{[i]}) + \nabla_1 K(\boldsymbol{\theta}_\tau^{[i]}, \boldsymbol{\theta})\,. \tag{3.73}
$$

The update for the ensemble of particles at level $\ell \in \{1, \ldots, L\}$ with step size $\delta > 0$ in the practical implementation then becomes

$$\boldsymbol{\theta}_{\tau+1}^{[i]} = \boldsymbol{\theta}_{\tau}^{[j]} - \delta \hat{\mathbf{g}}_{\tau}^{(\ell)}(\boldsymbol{\theta}_{\tau}^{[i]}), \quad i = 1, \ldots, N. \tag{3.74}$$

Computing the gradients (3.73) requires choosing a kernel K that defines the RKHS in which the gradients lie. As discussed in Section 3.1 a good choice of the kernel is necessary to ensure that the updates (3.3) and (3.6) minimize the KL divergence and converge to the desired target distribution as required by Assumptions 9 and 12. A common choice that works well in practice is the radial basis function kernel

$$K(\boldsymbol{\theta}, \boldsymbol{\theta}') = \exp\left(-\frac{1}{2\sigma} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2\right), \tag{3.75}$$

where the bandwidth σ controls the range of the interactions between the particles. For MLSVGD, as well as single-level SVGD, we must determine the number of iterations T_ℓ to perform before switching to the next level given the stopping criteria definition (3.42). This stopping criteria is not implementable because it requires monitoring the KL divergences $\mathrm{KL}\left(\mu_\tau \mid\mid \pi^{(\ell)}\right)$ in turn requires the normalized densities $\pi^{(\ell)}$ and the density of the MLSVGD approximation. Instead we monitor the average norm of the gradients

$$\bar{g}_{\tau}^{(\ell)} = \frac{1}{N} \sum_{i=1}^{N} \left\| \hat{\mathbf{g}}_{\tau}^{(\ell)}(\boldsymbol{\theta}_{\tau}^{[i]}) \right\|, \tag{3.76}$$

which approximates the expected norm of the gradient $\mathbb{E}_{\mu_\tau^{(\ell)}} \left\| \hat{\mathbf{g}}_{\tau}^{(\ell)}(\boldsymbol{\theta}_\tau) \right\|$. We terminate the SVGD updates (3.74) at level ℓ and switch to the next level whenever the average norm of the gradient (3.76) $\bar{g}_{\tau}^{(\ell)} \leq \epsilon_\ell$. In our implementation used for the numerical experiments, we find that setting $\epsilon_\ell = \epsilon$ works well in practice. This adaptive stopping criteria based on the gradient norm is motivated by [Duncan et al., 2019, Equation 61] which shows that for small

Algorithm 3: Discrete-time MLSVGD with approximate gradients and adaptive stopping criterion

Inputs: (unnormalized) densities $\pi^{(1)}, \dots, \pi^{(L)}$, initial particles $\{\boldsymbol{\theta}_0^{[i]}\}_{i=1}^N$, step size δ, tolerance ϵ;

Result: Particles $\{\boldsymbol{\theta}_\tau^{[i]}\}_{i=1}^N$

for $\ell = 1, \dots, L$ **do**

  Set $\tau = 0$;

  repeat

    Compute scores $s_i = \nabla \log \pi^{(\ell)}(\boldsymbol{\theta}_\tau^{[i]})$ for $i = 1, \dots, N$;

    for $i = 1, \dots, N$ **do**

      $\boldsymbol{\theta}_{\tau+1}^{[i]} = \boldsymbol{\theta}_\tau^{[i]} + \frac{\delta}{N} \left(\sum_{j=1}^N \nabla_1 K(\boldsymbol{\theta}_\tau^{[j]}, \boldsymbol{\theta}_\tau^{[i]}) + \sum_{j=1}^N K(\boldsymbol{\theta}_\tau^{[j]}, \boldsymbol{\theta}_\tau^{[i]}) s_j \right)$;

    end

    Estimate the norm of the gradient $\bar{g}_\tau^{(\ell)}$ as in (3.76) ;

    Set $\tau \leftarrow \tau + 1$;

  until $\bar{g}_\tau^{(\ell)} \leq \epsilon$;

end

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. Algorithm 12 outlines the practical implementation of MLSVGD that we use in the following numerical examples.

3.6 Numerical Experiments

The following two examples demonstrate the computational savings of MLSVGD over single-level SVGD for inference of a posterior in a Bayesian inverse problem depending on an underlying PDE. In both examples, we use a Matlab implementation with Intel Xeon CPU E5-2690 v2 processors, restricted to 8 cores and 32GB memory.

3.6.1 Nonlinear reaction diffusion

Set up

The first example we consider is inferring a nonlinear reaction term in a nonlinear reaction diffusion equation. Consider the spatial domain $\Omega = (0,1)^2$ and the nonlinear reaction diffusion equation

$$-\Delta u(x_1, x_2; \boldsymbol{\theta}) + q(u(x_1, x_2; \boldsymbol{\theta}), \boldsymbol{\theta}) = 100 \sin(2\pi x_1) \sin(2\pi x_2)\,, \quad \mathbf{x} = (x_1, x_2)^\top \in \Omega\,, \quad (3.77)$$

with homogeneous Dirichlet boundary conditions. The solution $u : \Omega \times \Theta \to \mathbb{R}$ depending on the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top \in \Theta = \mathbb{R}^2$ corresponds the concentration of a chemical undergoing a nonlinear reaction determined by the term

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

Let $F^{(\ell)} : \Theta \to C(\Omega)$ denote the forward models which map the parameters $\boldsymbol{\theta} \in \Theta$ to the numerical solution $u^{(\ell)}(\cdot; \boldsymbol{\theta})$. In particular, the model $F^{(\ell)}$ solves the PDE (3.77) by discretizing with finite differences on a grid with equidistant grid points and mesh width $h = 2^{-\ell-2}$. The resulting nonlinear system of equations is then solved to obtain a vector $\mathbf{u} \in \mathbb{R}^p$ ($p = (2^{\ell+2}-2)^2$) of the solution at the grid points by using Newton's method where the step length is determined by an inexact line search based on the Armijo condition [Nocedal and Wright, 2006]. The numerical solution $u^{(\ell)}$ is obtained as a piecewise linear interpolant between the grid points so that $u^{(\ell)}(\mathbf{x}_i; \boldsymbol{\theta}) = \mathbf{u}_i$ at the grid points $\mathbf{x}_i \in \Omega$ for $i = 1, \ldots, p$. Once the solution $u^{(\ell)}$ is obtained, the observation operator $\mathcal{B}^{\text{obs}} : C(\Omega) \to \mathbb{R}^{12}$ maps $u^{(\ell)}$ to its pointwise evaluations on the grid $[0.25i, 0.2j]$ for $i = 1, 2, 3$ and $j = 1, 2, 3, 4$. The full parameter-to-observable map is therefore, $G^\ell = \mathcal{B}^{\text{obs}} \circ F^{(\ell)}$ and we consider levels $\ell \in \{1, 2, 3\}$

(so $L = 3$). Note that because of the nonlinear dependence of the reaction term q on both the solution u and the parameter $\boldsymbol{\theta}$ we cannot write $F^{(\ell)}$ as a composition of an interpolation operator and a solution operator as we did for the examples 2.5.1 and 2.5.2 in Chapter 2. The observed data $\mathbf{y} \in \mathbb{R}^{12}$ is generated by evaluating a high-fidelity parameter-to-observable at the true parameters $\boldsymbol{\theta}^\star = (-\pi/4,\ 3)^\top$ and perturbed by Gaussian noise

$$\mathbf{y} = G^{(L+1)}(\boldsymbol{\theta}^\star) + \boldsymbol{\eta}\,.$$

The standard deviation of the added noise $\boldsymbol{\eta}$ is 0.5% of the true solution and is independent in each coordinate. Finally, the prior distribution π_0 is Gaussian $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ with

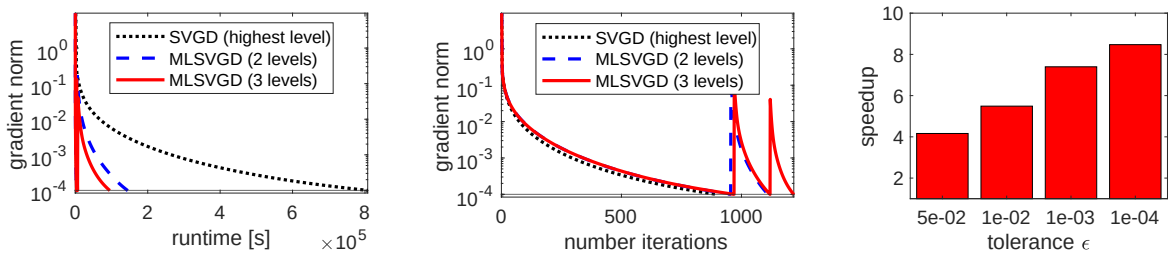
$$\boldsymbol{\mu}_0 = \begin{pmatrix} \pi/2 \\ 3/2 \end{pmatrix}, \quad \boldsymbol{\Sigma}_0 = \begin{bmatrix} 50 & 0 \\ 0 & 1/2 \end{bmatrix}.$$

SVGD and MLSVGD

For both SVGD and MLSVGD we use $N = 1000$ particles drawn from a Gaussian initial distribution $\mu_0 = N\left(\mathbf{1}_2,\ 10^{-4}\mathbf{I}_{2\times 2}\right)$. The kernel K that defines the RKHS for the gradient is taken to be the radial basis function kernel (3.75) with bandwidth $\sigma = 10^{-2}$. To compute the score functions $\nabla \log \pi^{(\ell)}$ of each posterior, we approximate the gradient with central finite differences with a mesh width of 2^{-6} and 12 total model evaluations per particle per iteration. In both SVGD and MLSVGD with a step size of $\delta = 10^{-1}$. The step size δ and kernel bandwidth σ are chosen manually to see ensure that SVGD performed at the highest level $L = 3$ converges. We run SVGD with respect to the high-fidelity level $\pi^{(3)}$ until the average norm of the gradients (3.76) falls below the tolerance ϵ, while for MLSVGD we consider both the levels $\ell \in \{1, 2, 3\}$ and $\ell \in \{1, 3\}$.

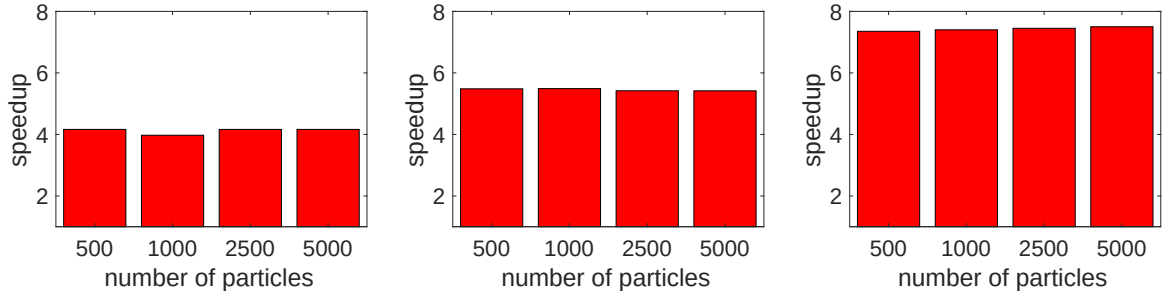
Runtime comparison of SVGD and MLSVGD

Figure 3.2 compares the average norm of the gradient (3.76) that defines the stopping criteria in Algorithm 12 over time across MLSVGD with 3 levels ($\ell \in \{1, 2, 3\}$), MLSVGD with 2 levels ($\ell \in \{1, 3\}$), and SVGD at the highest level $L = 3$ only when the tolerance is set to $\epsilon = 10^{-4}$. Figures 3.2(a,b) show that although MLSVGD requires more total iterations than SVGD, almost all of these iterations are on the lower levels which are significantly cheaper to perform. Because the surrogate densities $\pi^{(1)}$ and $\pi^{(2)}$ are good approximations to the high-fidelity density $\pi^{(3)}$, by the time MLSVGD switches to the final level the particles already have a favorable initialization. This leads to a factor of 8 speedup for MLSVGD with 3 levels over SVGD at the highest level only. MLSVGD with 2 levels ($\ell \in \{1, 3\}$) achieves a slightly lower speedup than MLSVGD with 3 levels in this example. Figure 3.2(c) shows the speedup of MLSVGD with 3 levels over SVGD at the highest level for different tolerances ϵ. As expected from the bounds on the cost complexities (3.16) and (3.13), we see that the speedup increases as $\epsilon \to 0$. Notice that for MLSVGD in Figures 3.2(a-b) there are spikes in the average gradient norm when the tolerance ϵ has been reached at level ℓ and MLSVGD switches to the next higher level. These spikes are due to the sudden change in the objective function $\mathrm{KL}\left(\cdot \,||\, \pi^{(\ell)}\right)$, which SVGD seeks to minimize, between levels and then shrink again quickly as the particles converge due to their good initialization from the previous level. The results from Figure 3.2(c) are extended in Figure 3.3 where we repeat this experiment and observe consistent speedups across for different numbers of particles $N \in \{500, 1000, 2500, 5000\}$. The speedup of MLSVGD over SVGD is consistent as the number of particles changes in this case because the costs of each iteration of MLSVGD and SVGD scale the same with respect to the number of particles.



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

Figure 3.2: Nonlinear reaction-diffusion: MLSVGD achieves speedups because most of the iterations are on lower, cheaper levels, in contrast to SVGD which performs all iterations on the highest, most expensive level. A spike in the gradient norm for MLSVGD indicates switching to a higher level.



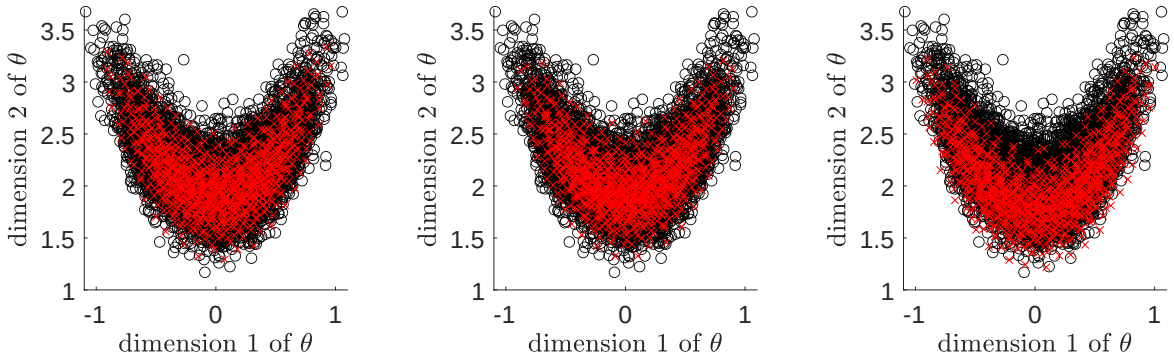
(a) tolerance $\epsilon = 5 \times 10^{-2}$ (b) tolerance $\epsilon = 10^{-2}$ (c) tolerance $\epsilon = 10^{-3}$

Figure 3.3: Nonlinear reaction-diffusion: The costs of both MLSVGD and SVGD scale quadratically with the number of particles due to the computation of the pairwise interactions through the kernel, which means that the speedups that MLSVGD obtains compared to SVGD in this example remain constant for different number of particles.

Accuracy comparison of SVGD and MLSVGD

Figure 3.4 shows the final samples (in red) obtained from running MLSVGD with 3 levels as well as the final samples attained from running SVGD at the lowest level $\ell = 1$ and at the highest level $L = 3$ when a tolerance of $\epsilon = 10^{-3}$ is set. The samples in black are 10^4 reference samples obtained by running the delayed-rejection adaptive Metropolis (DRAM) MCMC method [Haario et al., 2001, Haario et al., 2006] at the highest level. For the DRAM method a Gaussian proposal distribution is used and initialized with the covariance matrix $10^{-2}\mathbf{I}_{2\times 2}$. We take a burn-in period of 10^4 samples and then compute another 2×10^4 samples taking every other sample as a reference sample i.e. 10^4 total reference samples. From inspection of the samples in Figure 3.4 we see that MLSVGD and SVGD on the highest level (i.e. single-level SVGD) are indistinguishable and that when compared to the samples obtained by SVGD at the lowest level there are only minor differences visually. This is expected since SVGD at the highest level and MLSVGD converge asymptotically to the high-fidelity posterior $\pi^{(L)}$, while SVGD at the lowest level converges to surrogate density $\pi^{(1)}$. Although the samples obtained from SVGD at the lowest level are biased, they are distributed closely to the high-fidelity posterior and serve as a good initialization for higher levels as is done in MLSVGD. Figure 3.5(a) shows the error of the inferred posterior mean (i.e. mean of the particles) for each method compared to a ground truth reference value computed using the MCMC reference samples. For each method, the error of the particle mean is estimated by averaging over 10 independent runs

$$\frac{1}{10}\sum_{i=1}^{10}\left\|\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{(i)}\right\|_2,$$



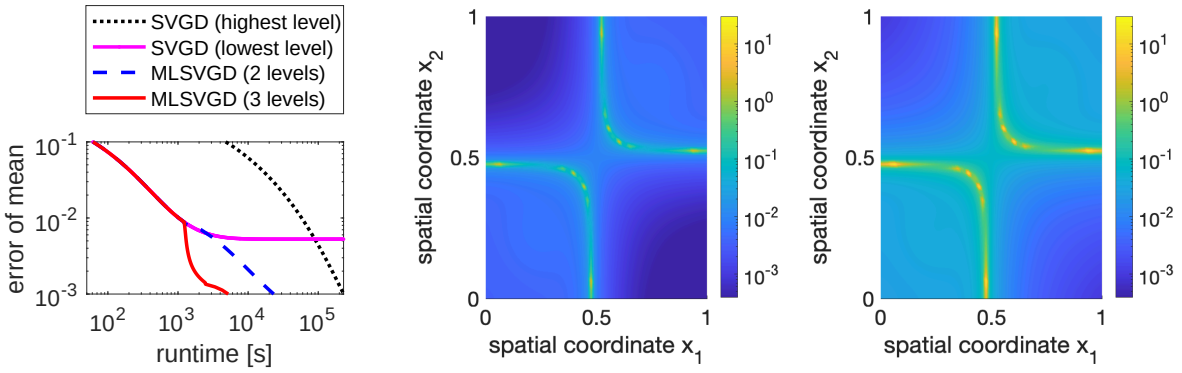
(a) MLSVGD (3-levels) (b) SVGD on highest level (c) SVGD on lowest level

Figure 3.4: Nonlinear reaction-diffusion: Sample particles ($N = 1000$) obtained with tolerance $\epsilon = 10^{-3}$ in red compared to the MCMC reference samples in black. (a) Samples obtained from MLSVGD with levels $\ell = 1, 2, 3$. (b) Samples obtained from running SVGD on the highest level $L = 3$. (c) Samples obtained from running SVGD on the lowest level $\ell = 1$.

where $\bar{\boldsymbol{\theta}}$ is the MCMC reference mean and $\hat{\boldsymbol{\theta}}^{(i)}$ is the i-th replicate of the particle mean

$$\hat{\boldsymbol{\theta}}^{(i)} = \frac{1}{N} \sum_{j=1}^{N} \boldsymbol{\theta}^{[j]},$$

with samples $\{\boldsymbol{\theta}^{[j]}\}_{j=1}^{N}$ obtained from the i-th run of each method. We see immediately that SVGD at the lowest level is biased as the error of the inferred posterior mean levels off despite increasing computational costs (runtime). On the other hand, SVGD at the highest level correctly infers the posterior mean but is computationally expensive with MLSVGD achieving more than one order of magnitude in speedup for the same error. From this perspective, MLSVGD leverages the computational savings of SVGD on the lowest level while retaining the accuracy guarantees that SVGD on the highest level enjoys. Figures 3.5(b, c) show the pointwise error of the finite-difference solution u of (3.77) 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 computed at the MCMC reference mean. For the same costs as SVGD, MLSVGD recovers a more accurate solution with lower pointwise error.



(a) error w.r.t. MCMC reference (b) MLSVGD (c) SVGD (same costs as (b))

Figure 3.5: Nonlinear reaction-diffusion: MLSVGD infers the posterior mean with error 10^{-3} with respect to an MCMC reference with more than one order of magnitude speedup compared to SVGD.

3.6.2 Euler-Bernoulli beam

In this example we consider the same problem as in Section 2.5.2 of inferring the effective stiffness of an Euler-Bernoulli beam.

Set up

The set up of the inverse problem is the similar to the set up presented in Section 2.5.2, and so we only highlight the differences here. The high-fidelity model is now taken to use 601 grid points in the finite difference discretization of the differential equation (2.58). Additionally, we consider five surrogate models ,so that $L = 6$, which discretize (2.58) using 51, 101, 201, 301, 401, and 501 grid points, respectively. The observation operator $\mathcal{B}^{\text{obs}} : C[0,1] \to \mathbb{R}^{41}$ evaluates the finite difference solution $u^{(\ell)}$ at 41 equally spaced points $x_i = (i-1)/40$ for $i = 1, \ldots, 41$ throughout the domain $\Omega = [0,1]$. Note that here we also observe the solution at the left end-point which is fixed by the boundary conditions. Observational data $\mathbf{y}$ is generated by computing the finite-difference solution to (2.58) with 601 grid points and constant effective stiffness $E(x) = 1$, and then perturbing with the output of the observation operator with mean-zero 0.01% Gaussian noise. In this example, we

consider different dimensions d of the problem: $d \in \{3, 6, 9, 12, 16\}$ as opposed to Section 2.5.2 where we only considered $d = 6$. For this we take the prior to be log-normal with mean $\boldsymbol{\mu}_0 = \mathbf{1}_d$ and covariance $\boldsymbol{\Sigma}_0 = 5 \times 10^{-2}\mathbf{I}_{d\times d}$.

Results for SVGD and MLSVGD

We compare the results for MLSVGD with 6 levels ($\ell \in \{1, \ldots, 6\}$), MLSVGD with 3 levels ($\ell \in \{1, 3, 6\}$), and SVGD at the highest level $L = 6$. For all methods we sample $N = 500$ particles from the initial distribution $\mu_0 = N(\mathbf{1}_d,\ 4\times 10^{-4}\mathbf{I}_{d\times d})$. As in the nonlinear reaction-diffusion example we determine the step sizes for each dimension manually. For $d = 3$ we set $\delta = 10^{-3}$, for $d = 6, 9$ we set $\delta = 10^{-2}$, and for $d = 12, 16$ we set $\delta = 5 \times 10^{-3}$. Similarly for $d = 3$ the kernel bandwidth is chosen to be $\sigma = 10^{-6}$, for $d = 6, 9$ it is $\sigma = 10^{-5}$, and for $d = 12, 16$ it is $\sigma = 5 \times 10^{-5}$. Figure 3.6 shows the convergence of MLSVGD and SVGD across dimensions $d \in \{3, 6, 9, 12, 16\}$ as well as the error of the particle means with respect to the MCMC reference mean when the tolerance $\epsilon = 5 \times 10^{-3}$. Again we compute the MCMC reference mean as in the nonlinear reaction-diffusion example by averaging over 10 replicates. We see that each dimension shares the same behavior, namely that MLSVGD requires more iterations for convergence, most of which occur on the lowest level leading to runtime improvements over SVGD. Note that MLSVGD with 3 levels achieves about the same speedup as MLSVGD with 6 levels, indicating that adding more intermediate levels has diminishing returns in terms of computational savings and cannot further reduce the costs. By looking at the last column of plots in Figure 3.6 we see that MLSVGD recovers the true particle mean faster than SVGD with the same accuracy. The speedups of MLSVGD with 3 levels over SVGD for each dimension are summarized in Figure 3.7 where the speedup is consistently between a factor of 6 to 10 regardless of the dimension. For each dimension, Figure 3.8 shows the relative pointwise error of the finite-difference solution u of (2.58) computed at the final particles obtained with MLSVGD and single-level SVGD

when compared to the MCMC reference solution. In particular, Figure 3.8 shows error bars corresponding to minimum and maximum pointwise relative error over each final particle in the ensemble. The results show that MLSVGD achieves a similar error as single-level SVGD despite MLSVGD being faster by up to one order of magnitude as shown in Figure 3.6. Moreover, the variance of the errors are comparable as well as indicated by the minimum and maximum error bars.

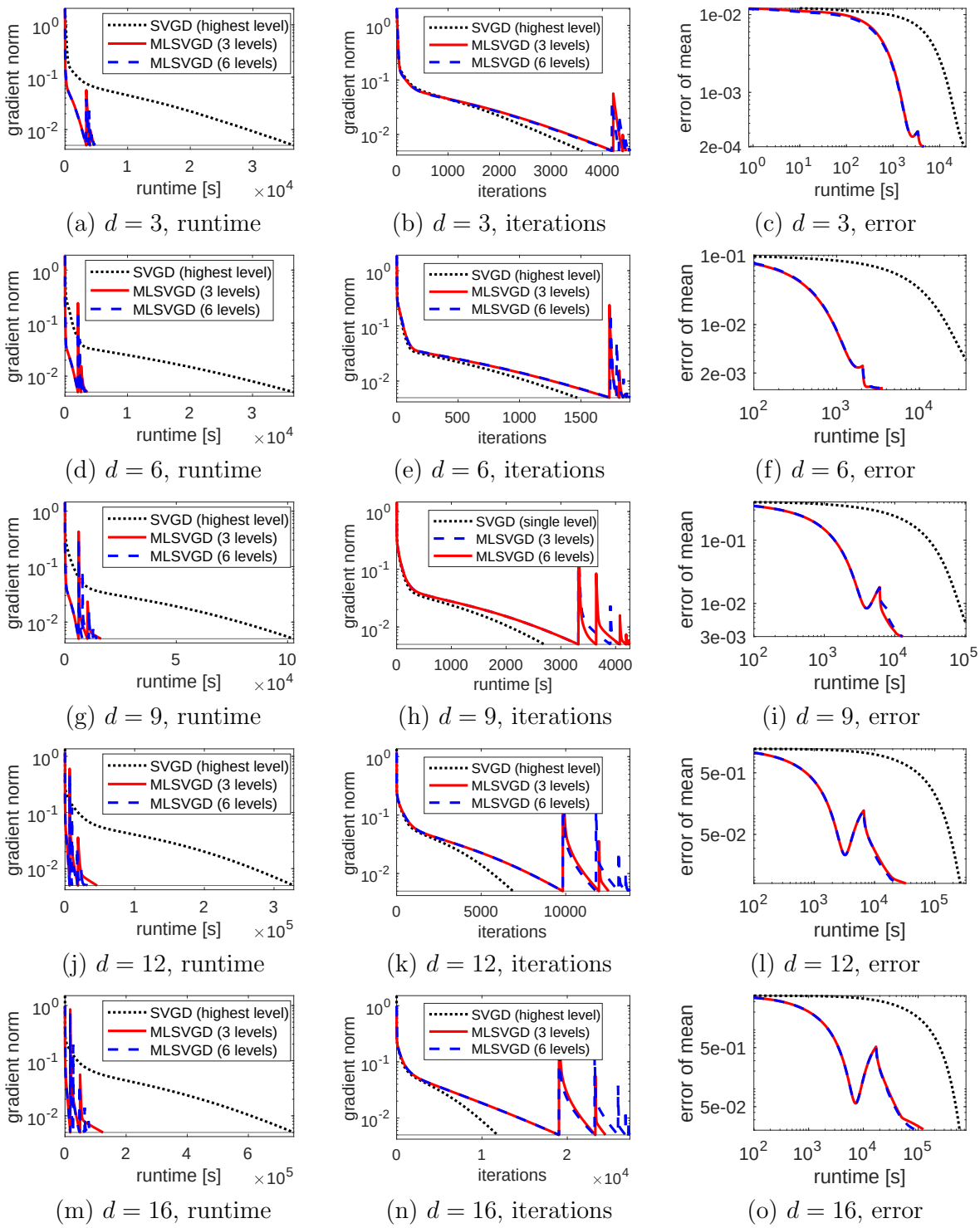
3.7 Inferring ice sheet flow of the Arolla glacier

In this section we demonstrate that MLSVGD outperforms SVGD for the task of inferring the basal sliding coefficient to understand the flow of a glacial mass sliding across underlying bedrock based on pointwise velocity observations. This problem is an ISMIP-HOM [Pattyn et al., 2008] benchmark problem and is made difficult by the nonlinearity in the flow of the ice, the curvature of the domain, and the computational cost of solving the model at a fine resolution. For the set up of the Bayesian inverse problem for inferring the basal sliding coefficient, we closely follow the problem formulation introduced in [Petra et al., 2014]. We use a Python 3 implementation with FEniCS and hIPPYlib [Alnaes et al., 2015, Villa et al., 2016, Villa et al., 2018, Villa et al., 2021] to discretize the forward model and perform adjoint-based gradient computations required by SVGD. All runtimes were measured on Intel Xeon Platinum 8268 24C 205W 2.9GHz Processors and the computation of the gradients $\nabla \log \pi^{(\ell)}(\boldsymbol{\theta}_\tau^{[j]})$ at each iteration was parallelized over 32 cores.

3.7.1 Forward model of sliding of Arolla glacier ice

Following [Petra et al., 2014], the glacier is modeled as a sliding mass of ice whose velocity is determined primarily by the force of gravity and a frictional force experienced while sliding against the underlying bedrock. The ice itself is modeled as a non-Newtonian, viscous, and incompressible fluid whose velocity field $\boldsymbol{u} : \Omega \to \mathbb{R}^2$ over the glacier domain $\Omega \subset \mathbb{R}^2$, shown in Figure 3.9, is the solution to the Stokes equation

$$\begin{aligned} \nabla \cdot \boldsymbol{u} &= 0, \quad \text{in } \Omega\,, \\ -\nabla \cdot \boldsymbol{\sigma}_u &= \rho \boldsymbol{g}, \quad \text{in } \Omega\,, \end{aligned} \tag{3.78}$$



(a) $d = 3$, runtime (b) $d = 3$, iterations (c) $d = 3$, error

(d) $d = 6$, runtime (e) $d = 6$, iterations (f) $d = 6$, error

(g) $d = 9$, runtime (h) $d = 9$, iterations (i) $d = 9$, error

(j) $d = 12$, runtime (k) $d = 12$, iterations (l) $d = 12$, error

(m) $d = 16$, runtime (n) $d = 16$, iterations (o) $d = 16$, error

Figure 3.6: Euler-Bernoulli: Runtime, iterations, and error with respect to MCMC reference of MLSVGD and SVGD for dimensions $d \in \{3, 6, 9, 12, 16\}$.

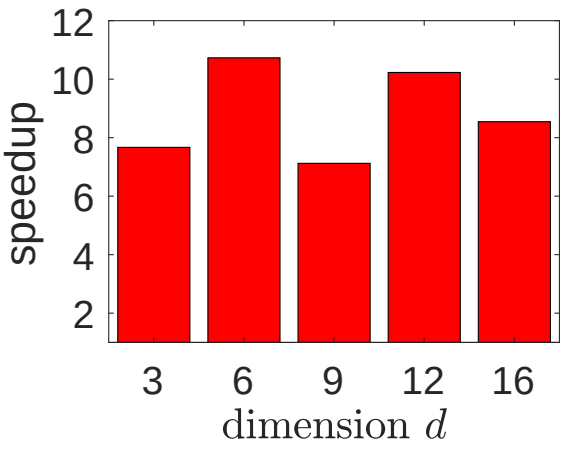
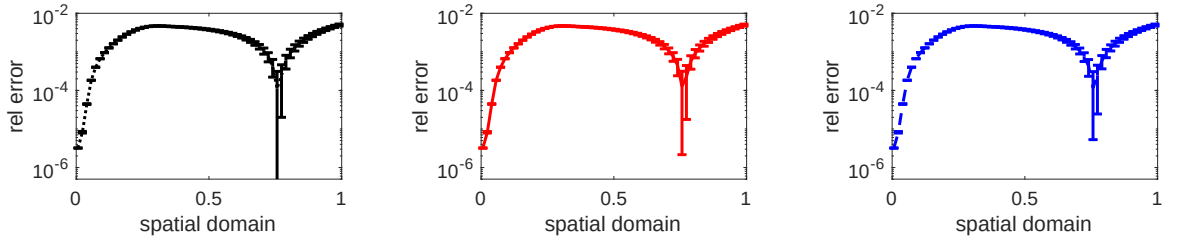


Figure 3.7: Euler-Bernoulli: MLSVGD achieves speedups between 6–10 across different dimensions in this example compared to SVGD.

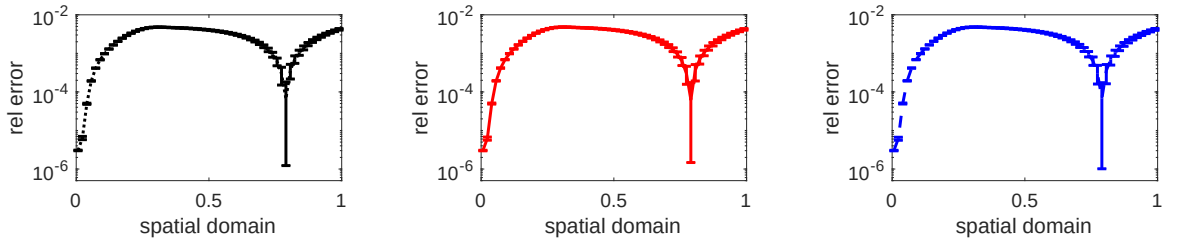
with stress tensor $\boldsymbol{\sigma}_u$, ice density $\rho = 910$ [kg/m^3], and the downwards gravitational force is $\boldsymbol{g} = (0, -9.81)$ [m/s^2]. The boundary conditions along the top Γ_t and bottom Γ_b of the glacier where the ice slides across the bedrock (see Figure 3.9) are given as

$$\begin{aligned} \boldsymbol{n}^\top (\boldsymbol{\sigma}_u \boldsymbol{n} + \lambda \boldsymbol{u}) &= 0, \quad \text{on } \Gamma_b\,, \\ \mathbf{T}\boldsymbol{\sigma}_u \boldsymbol{n} + \exp(\beta)\mathbf{T}\boldsymbol{u} &= \mathbf{0}, \quad \text{on } \Gamma_b\,, \\ \boldsymbol{\sigma}_u \boldsymbol{n} &= \mathbf{0}, \quad \text{on } \Gamma_t\,. \end{aligned} \tag{3.79}$$

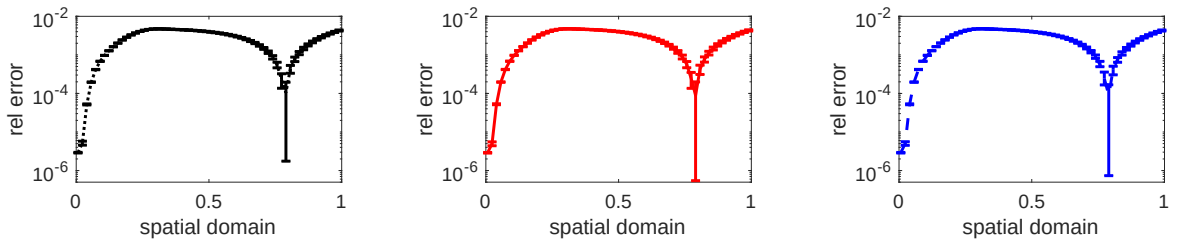
The vector $\boldsymbol{n}$ represents the outward unit normal vector and $\mathbf{T} = \mathbf{I} - \boldsymbol{n}\boldsymbol{n}^\top$ is the tangential projection. The first boundary condition where $\boldsymbol{n}^\top (\boldsymbol{\sigma}_u \boldsymbol{n} + \lambda \boldsymbol{u}) = 0$ approximates a no outflow condition $\boldsymbol{u} \cdot \boldsymbol{n} = 0$, which is difficult to enforce directly due to the curvature of the domain Ω. In this example, we set $\lambda = 10^6$. The second boundary condition involves the log basal sliding coefficient field $\beta : [0, 5000] \to \mathbb{R}$ which determines the frictional force by relating the tangential velocity to the tangential traction as the ice sheet slides downhill



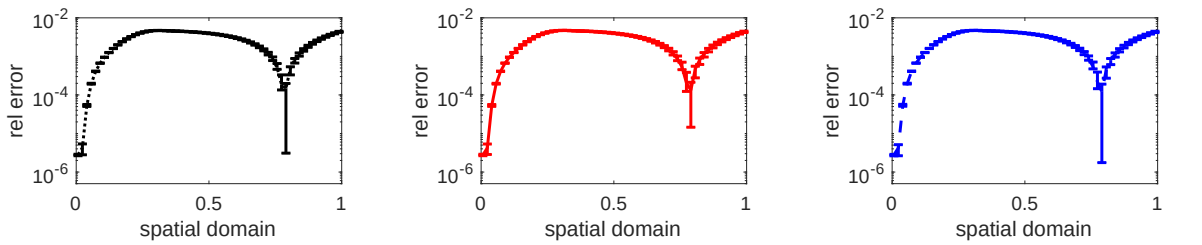
(a) $d = 3$, single-level SVGD (b) $d = 3$, MLSVGD (3 levels) (c) $d = 3$, MLSVGD (6 levels)



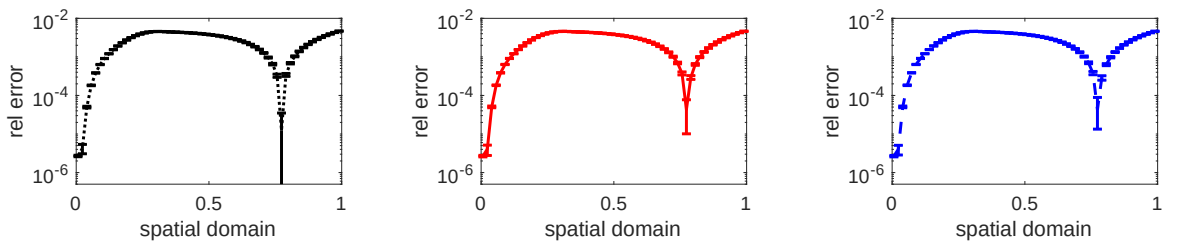
(d) $d = 6$, single-level SVGD (e) $d = 6$, MLSVGD (3 levels) (f) $d = 6$, MLSVGD (6 levels)



(g) $d = 9$, single-level SVGD (h) $d = 9$, MLSVGD (3 levels) (i) $d = 9$, MLSVGD (6 levels)



(j) $d = 12$, single-level SVGD (k) $d = 12$, MLSVGD (3 levels) (l) $d = 12$, MLSVGD (6 levels)



(m) $d = 16$, single-level SVGD (n) $d = 16$, MLSVGD (3 levels) (o) $d = 16$, MLSVGD (6 levels)

Figure 3.8: Euler-Bernoulli: Minimum and maximum of pointwise error over the ensemble of inferred solutions for $d \in \{3, 6, 9, 12, 16\}$.

across the underlying bedrock. The stress tensor can be split into

$$\boldsymbol{\sigma}_u = \boldsymbol{\tau}_u - \mathbf{I}p,$$

with pressure p and deviatoric stress tensor

$$\boldsymbol{\tau}_u = 2\eta(\mathbf{u})\dot{\boldsymbol{\varepsilon}}(\boldsymbol{u})\,,$$

defined in terms of the strain rate tensor

$$\dot{\boldsymbol{\varepsilon}}(\boldsymbol{u}) = \frac{1}{2}\left(\nabla\mathbf{u} + \nabla\mathbf{u}^\top\right)\,,$$

and the effective viscosity

$$\eta(\boldsymbol{u}) = \frac{1}{2}A^{-\frac{1}{n}}\dot{\boldsymbol{\varepsilon}}_{\text{II}}^{\frac{1-n}{2n}}\,.$$

The constants that determine the effective viscosity are Glen's flow law exponent $n = 3$ and the flow rate factor $A = 10^{-16}$ [$\text{Pa}^{-n}\text{a}^{-1}$] (Pascals and years, respectively) and the second invariant is

$$\dot{\boldsymbol{\varepsilon}}_{\text{II}} = \frac{1}{2}\text{Tr}\left(\dot{\boldsymbol{\varepsilon}}_{\mathbf{u}}^2\right)\,,$$

where Tr denotes the trace operator. To solve (3.78), we discretize (3.78)–(3.79) using Taylor-Hood finite elements on a triangular mesh where the velocity is discretized with quadratic Lagrange elements and the pressure is discretized with linear Lagrange elements. The resulting nonlinear system is then solved using a constrained Newton solver with the tolerance set to 10^{-6} on the L^2 norm of the gradient. In this example, we consider three different refinements of the mesh giving rise to the high-fidelity model and two low-fidelity models. The high-fidelity forward model $F^{(3)}$ ($L = 3$) approximates the solution operator that maps the log basal sliding coefficient field β to the velocity field solution $\boldsymbol{u}$ using

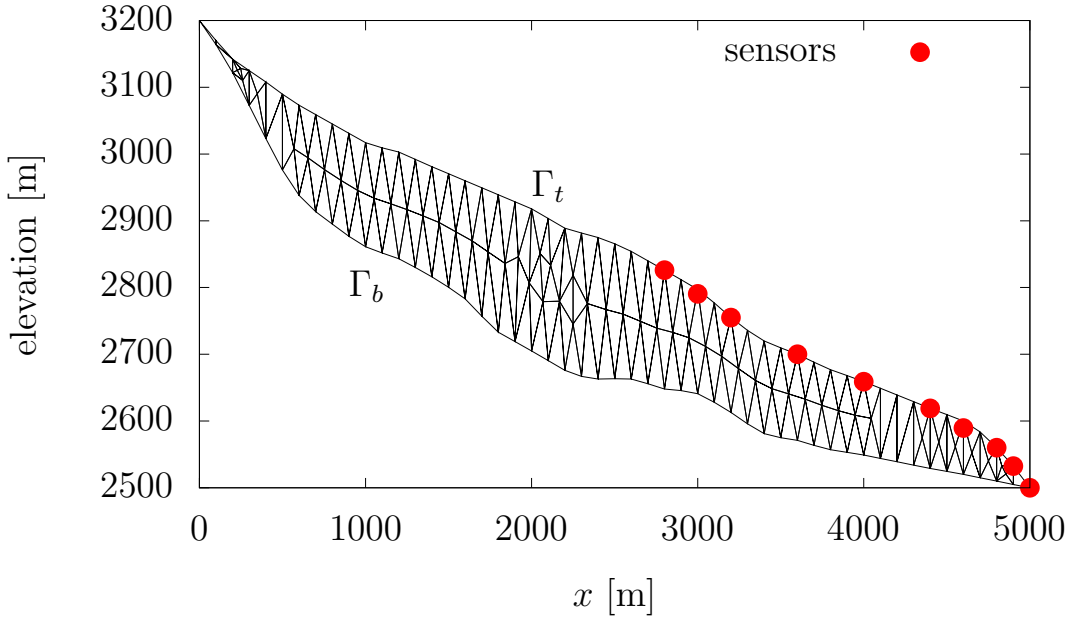


Figure 3.9: The Arolla domain Ω along with the location of the sensors (red) where velocity measurements are taken.

3,602 and 501 degrees of freedom for the velocity and pressure components, respectively. Similarly, the coarsest low-fidelity model $F^{(1)}$ uses 448 and 73 degrees of freedom for the velocity and pressure and the second low-fidelity model $F^{(2)}$ uses 1002 and 151 degrees of freedom, respectively.

3.7.2 Setup of Bayesian inverse problem

For this problem, we are interested in inferring the log basal sliding coefficient field β, which through the boundary condition along the bottom of the domain Γ_b, determines the velocity of the ice as it slides along the bedrock. We approximate the coefficient field in a finite dimensional space, in order to be compatible with the SVGD algorithm (c.f. Section 3.1), by discretizing the coefficient field $\beta : [0, 5000] \to \mathbb{R}$ with a vector $\boldsymbol{\beta} \in \mathbb{R}^d$ ($d = 25$) that we

aim to infer from the observational data. The parameter vector $\boldsymbol{\beta} \in \mathbb{R}^{25}$ corresponds to 25 equally-spaced pointwise evaluations of the coefficient field β throughout the domain $[0, 5000]$ and the finite dimensional approximation arises by considering piecewise linear interpolants through these points. In particular, let $\mathcal{I}^{\text{int}}$ be the interpolation operator that maps a vector $\boldsymbol{\beta} \in \mathbb{R}^{25}$ to its piecewise linear interpolant $\bar{\beta} : [0, 5000] \to \mathbb{R}$ defined at the nodes $x_i = 5000(i-1)/24$ by $\bar{\beta}(x_i) = \boldsymbol{\beta}_i$ for $i = 1, \ldots, 25$. Given the piecewise linear interpolant $\bar{\beta}$, the forward models $F^{(\ell)}$ for $\ell = 1, 2, 3$ map the parameter to the corresponding velocity field $\boldsymbol{u}$ that solves (3.78) with the boundary conditions (3.79). Finally, define the observation operator $\mathcal{B}^{\text{obs}}$ which maps the solution $\boldsymbol{u}$ of (3.78), given by the output of the forward models $F^{(\ell)}$, to a 20 dimensional vector of horizontal and vertical velocity measurements at 10 sensor locations throughout the right side of the domain along the surface of the glacier as shown in 3.9. The full parameter-to-observable map $G^{(\ell)} : \mathbb{R}^{25} \to \mathbb{R}^{20}$ is therefore

$$G^{(\ell)} = \mathcal{B}^{\text{obs}} \circ F^{(\ell)} \circ \mathcal{I}^{\text{int}} ,$$

for $\ell = 1, 2, 3$. We generate synthetic observational velocity data $\mathbf{y} \in \mathbb{R}^{20}$

$$\mathbf{y} = G^{(L+1)}(\boldsymbol{\beta}^*) + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim N(\mathbf{0}, \Gamma) ,$$

by solving (3.78) using a further refinement of the high-fidelity mesh (denoted by level $L+1$) and then corrupting with Gaussian noise. The noise covariance matrix Γ is diagonal with $\sigma_{\text{vert}} = 3$ and $\sigma_{\text{horz}} = 18$ corresponding to the vertical and horizontal velocity measurements, respectively. The true parameter $\boldsymbol{\beta}^\star \in \mathbb{R}^{25}$ used to generate the data $\mathbf{y}$ is obtained by taking pointwise evaluations $\boldsymbol{\beta}_i^\star = \beta_0(x_i)$ for $x_i = 5000(i-1)/24$ for $i = 1, \ldots, 25$ of the true

coefficient field

$$\beta_0(x) = \log \begin{cases} 1000 + 1000 \sin\left(\frac{3\pi x}{5000}\right) + \delta & \text{if } 0 \leq x < 2500, \\ 1000\left(16 - \frac{x}{250}\right) + \delta & \text{if } 2500 \leq x < 4000, \\ 1000 + \delta & \text{if } 4000 \leq x < 5000\,, \end{cases}$$

and $\delta = 10^{-6}$ is a small positive constant to ensure that the log basal coefficient field remains bounded. The prior π_0 is Gaussian with mean perturbed from the true parameters $\boldsymbol{\beta}^\star$ and a diagonal covariance matrix $\boldsymbol{\Sigma}_0 = 5 \times 10^{-2}\mathbf{I}_{25\times 25}$.

3.7.3 Numerical results

In the following we compare the performance of MLSVGD and SVGD. We run both SVGD and MLSVGD with $N = 1,000$ particles from the 25-dimensional standard normal distribution, a step size of $\delta = 5 \times 10^{-2}$, and the Gaussian radial basis function kernel (3.75) with the bandwidth parameter $\sigma = 10^{-1}$. The bandwidth parameter is kept constant, but is comparable to the one obtained from using the median heuristic presented in [Liu and Wang, 2016]. The gradients of the log posterior density needed for 12 are computed using adjoints with hIPPYlib [Villa et al., 2016, Villa et al., 2018, Villa et al., 2021] similar to the advection-diffusion example in Section 2.5.3. The quantity of interest is the high-fidelity posterior mean $\mathbb{E}_{\pi^{(L)}}[\boldsymbol{\beta}]$. A reference value $\hat{\boldsymbol{\beta}}^{\text{Ref}}$ of $\mathbb{E}_{\pi^{(L)}}[\boldsymbol{\beta}]$ is computed using the preconditioned Crank-Nicolson (pCN) method [Cotter et al., 2013] where we run 100 independent chains and use a burn-in period of 10,000 samples for each chain to obtain 10^7 total samples. The parameter in the pCN algorithm is set to 10^{-2}.

Number of iterations and runtime of SVGD and MLSVGD

As in the nonlinear reaction-diffusion 3.6.1 and Euler-Bernoulli 3.6.2 examples, we assess the convergence of SVGD and MLSVGD by monitoring the norm of the gradient at each iteration. With a gradient tolerance of $\epsilon = 10^{-2}$, MLSVGD achieves a speedup of a factor of 5 over SVGD despite requiring more iterations as shown in Figures 3.10(a,c). Similar to both of the examples in Section 3.6, over 80% of the iterations for MLSVGD are performed on the lowest level $\ell = 1$ with the coarsest mesh. Because the low-fidelity model is orders of magnitude faster than the high-fidelity model in this case, MLSVGD is able to quickly (in terms of runtime) converge to the low-fidelity posterior $\pi^{(1)}$, which serves as a good initial distribution for the following two levels. On the other hand, SVGD requires a comparable number of iterations, all of which are at the high-fidelity level, resulting in high computational costs. Figure 3.10(b,d) shows that both SVGD and MLSVGD accurately infer the quantity of interest (posterior mean) in terms of the relative error

$$\text{rel}(\boldsymbol{\beta}) = \frac{\left\| \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\text{Ref}} \right\|_2}{\left\| \hat{\boldsymbol{\beta}}^{\text{Ref}} \right\|_2} \tag{3.80}$$

when compared to the MCMC reference value and suggests that the distribution of the particles $\{\boldsymbol{\theta}_\tau^{[j]}\}_{j=1}^N$ converges to the high-fidelity target posterior $\pi^{(L)}$.

Speedups

Because MLSVGD leverages the low-fidelity models to find a good initialization of the particles before the final level, it is able to accurately infer the high-fidelity posterior mean $\mathbb{E}_{\pi^{(L)}}[\boldsymbol{\beta}]$ is less than a quarter of the time that SVGD takes. Snapshots of the particle mean at different fixed amounts of training time (runtime) for both MLSVGD and SVGD are shown in Figure 3.11. The left column shows snapshots of the MLSVGD inferred parameter

mean (red) at different times while the right column shows snapshots of the SVGD inferred parameter mean (blue) at the same times. In each plot the solid light gray curve shows the computed reference value. We see that after only 2 hours MLSVGD is able to recover the posterior mean, and hence the approximation to the coefficient field β, whereas SVGD still has not even after 8 hours. Notice that the coordinates $i = 12, \ldots, 25$ of the parameter vector $\boldsymbol{\beta}$, which correspond to the coefficient field on the right side of the domain $[2500, 5000]$, are recovered much faster due to the location of the sensors that observe the velocity as shown in Figure 3.9. Figure 3.12 shows the inferred velocity field $\boldsymbol{u}$ by solving (3.78) with the inferred parameter mean after approximately 8 hours of run time over 32 cores. We see that the velocity field obtained with the MLSVGD inferred parameter mean closely matches the velocity field obtained with the ground truth reference value of the mean while SVGD fails to recover the correct velocity field within the same amount of time. Again we see that the inferred velocity on the left side of the domain $[0, 2500]$ is inaccurate concurrent with the slow recovery of the parameter in this region. Moreover, note that the magnitude of the velocity is overestimated for SVGD which is consistent with the fact that the parameter, which controls the frictional forces to resist the downward pull of gravity, is underestimated.

Sample quality

One of the advantages of SVGD is that, with an appropriately chosen kernel, the samples may be more evenly spread out due to the repulsive interaction between particles. These repulsive interactions prevent the particles from clustering near each other allowing SVGD to avoid high correlations that some MCMC methods suffer from resulting in slow convergence. One way that sample quality can be measured is with the maximum mean discrepancy (MMD) [Gretton et al., 2012]

$$\mathrm{MMD}[\mu, \nu]^2 = \sup_{\|f\|_{\mathcal{H}} \leq 1} \left(\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]\right)^2 , \tag{3.81}$$

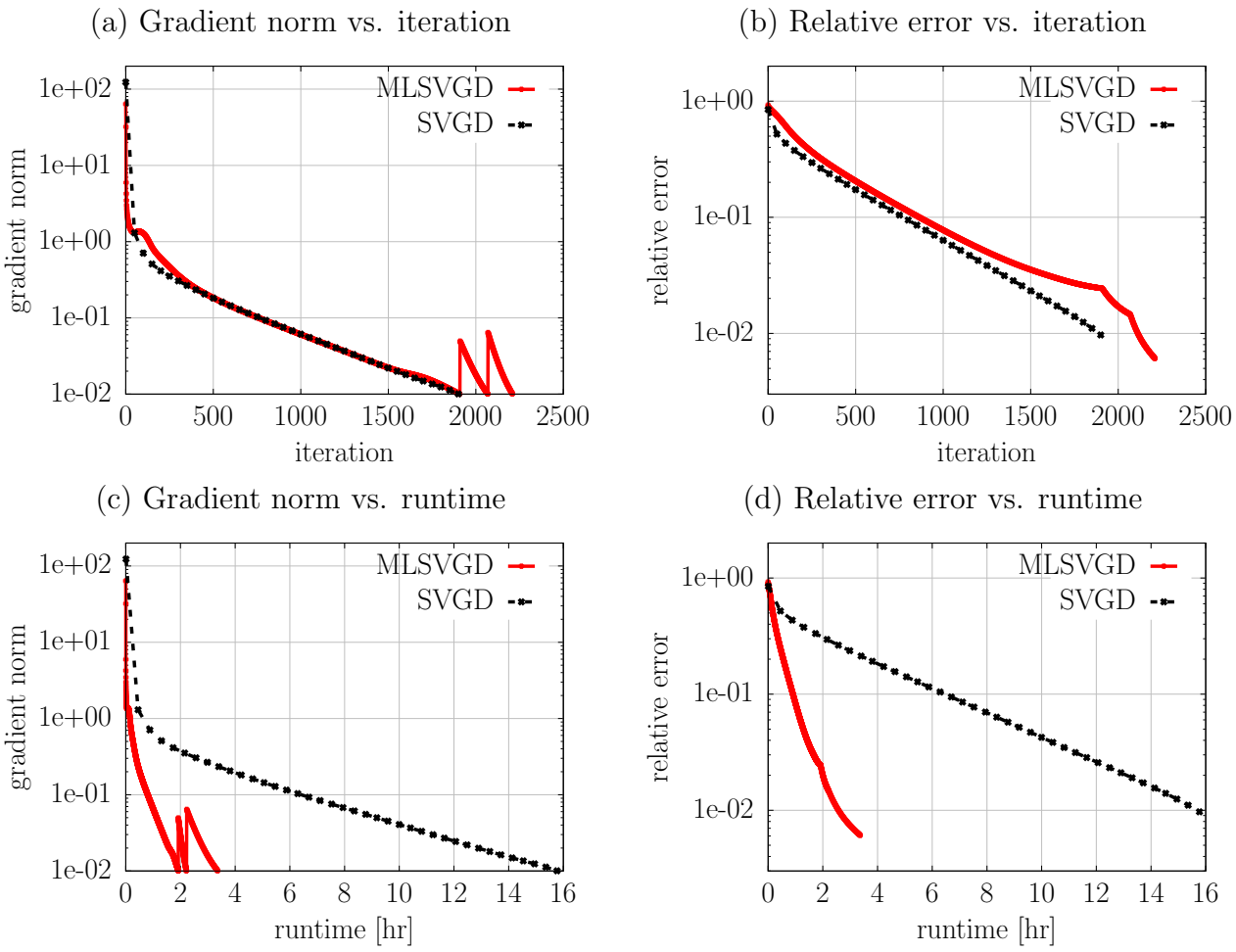
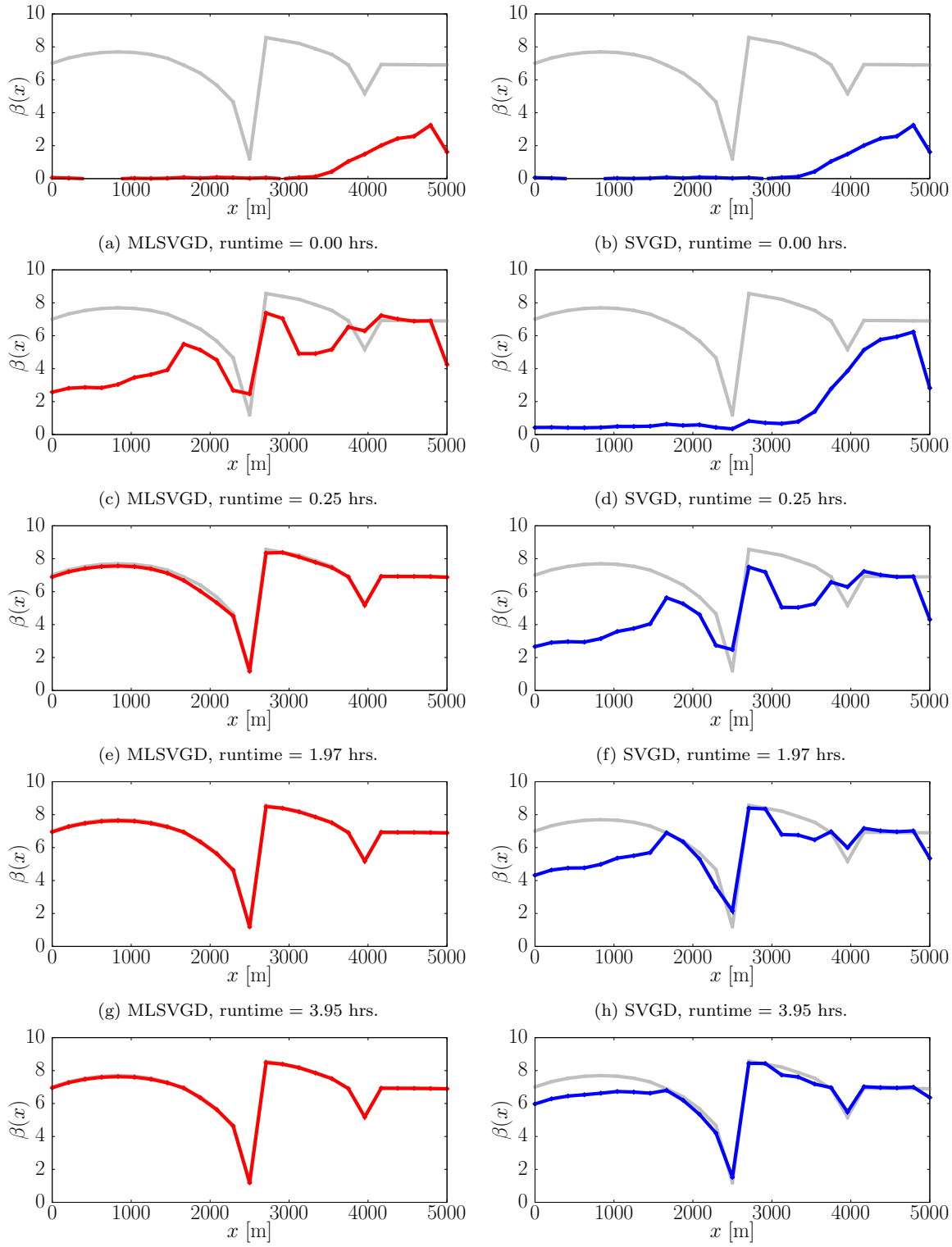


Figure 3.10: **(a)** The average gradient norm $\bar{g}_\tau$ vs. iteration for MLSVGD and SVGD with a tolerance of $\epsilon = 10^{-2}$. **(b)** The relative error of MLSVGD and SVGD compared to an MCMC reference vs. iteration. **(c)** The average gradient norms vs. the actual runtime in hours over 32 cores. **(d)** The relative error vs. actual runtime.



(a) MLSVGD, runtime = 0.00 hrs.

(b) SVGD, runtime = 0.00 hrs.

(c) MLSVGD, runtime = 0.25 hrs.

(d) SVGD, runtime = 0.25 hrs.

(e) MLSVGD, runtime = 1.97 hrs.

(f) SVGD, runtime = 1.97 hrs.

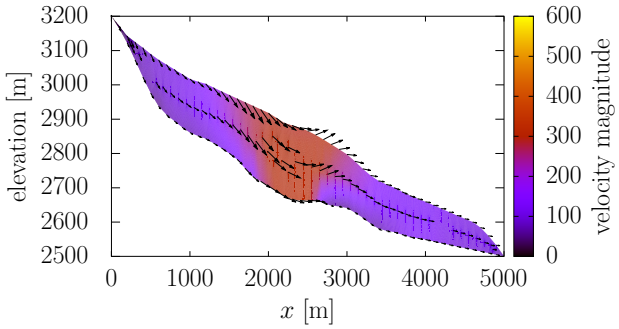
(g) MLSVGD, runtime = 3.95 hrs.

(h) SVGD, runtime = 3.95 hrs.

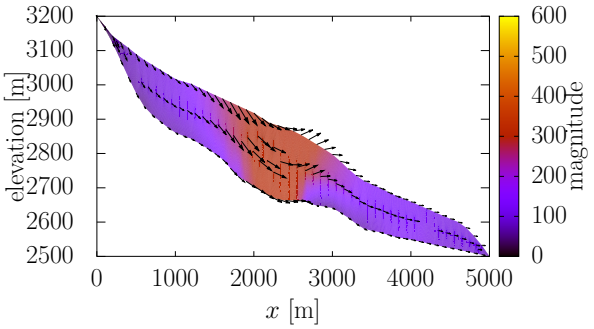
(i) MLSVGD, runtime = 7.90 hrs.

(j) SVGD, runtime = 7.90 hrs.

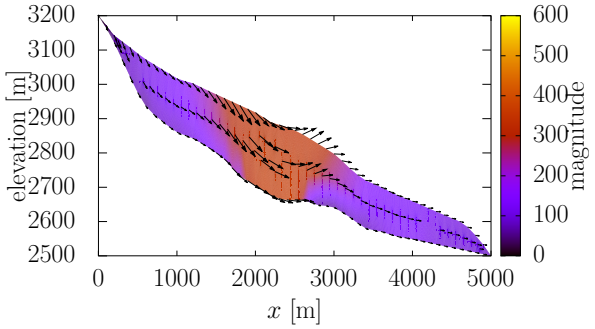
Figure 3.11: Parameter snapshots of MLSVGD (**Left**, red) and SVGD (**Right**, blue).



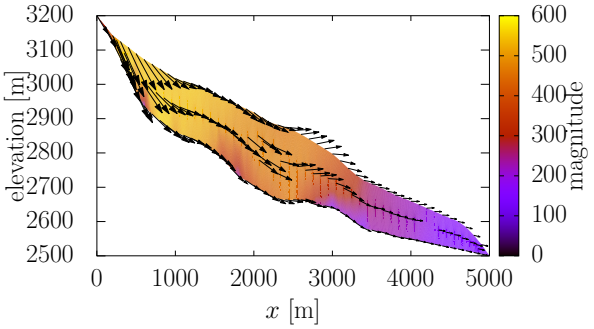
(a) True velocity field



(b) Reference velocity field



(c) MLSVGD inferred velocity field



(d) SVGD inferred velocity field

Figure 3.12: (**a**) The true velocity field given by $\boldsymbol{\beta}^\star$. The color indicates the magnitude of the velocity in [m a^{-1}] (meters per year). (**b**) The reference velocity field computed using $\hat{\boldsymbol{\beta}}^{\text{Ref}}$ of the posterior mean. (**c**) The velocity field corresponding to the inferred parameters using MLSVGD after eight hours. (**d**) The velocity field corresponding to the inferred parameters using SVGD with equivalent costs as MLSVGD (eight hours of runtime).

where $\mathcal{H}$ is the RKHS with kernel K and is zero if and only if the distributions are equal $\mu = \nu$. When the expectations in (3.81) cannot be evaluated exactly, one may use the following estimator [Gretton et al., 2012, Eq. 5], which leverages the RKHS structure, instead

$$\widehat{\mathrm{MMD}}\left(\{\mathbf{x}_i\}_{i=1}^N, \{\mathbf{y}_j\}_{j=1}^M\right)^2 = \frac{1}{N^2}\sum_{i=1}^N\sum_{i'=1}^N K(\mathbf{x}_i, \mathbf{x}_{i'}) + \frac{1}{M^2}\sum_{j=1}^M\sum_{j'=1}^M K(\mathbf{y}_j, \mathbf{y}_{j'}) - \frac{2}{NM}\sum_{i=1}^N\sum_{j=1}^M K(\mathbf{x}_i, \mathbf{y}_j)\,, \quad (3.82)$$

but requires samples $\{\mathbf{x}_i\}_{i=1}^N \sim \mu$ and $\{\mathbf{y}_j\}_{j=1}^M \sim \nu$ from both distributions. In our setting, we approximate samples from the target distribution by using pCN with $\beta = 0.01$. We run pCN for a burn-in period of 20,000 iterations and then an additional 100,000 iterations taking every 5th sample for $M = 20,000$ samples in total. Note that we only choose to take every 5th sample to reduce the computational cost of computing the MMD estimator (3.82), which scales quadratically in the number of samples. These 20,000 samples serve as reference samples from the high-fidelity target distribution $\pi^{(L)}$. Figure 3.13 below shows the estimated squared MMD for MLSVGD, SVGD, and MCMC (pCN) with equal sample sizes ($N = 1,000$). Both MLSVGD and SVGD have comparable sample quality and outperform the quality of samples produced by MCMC, which has a higher MMD. When computing the MMD for MCMC, the samples were taken from a chain that was independent of the one used to generate the 20,000 reference samples. For this independent chain we again used an initial burn-in period of 20,000 iterations and then take the following $1,000$ samples. Note that for this second chain we do not skip samples so as not to ignore their autocorrelation.

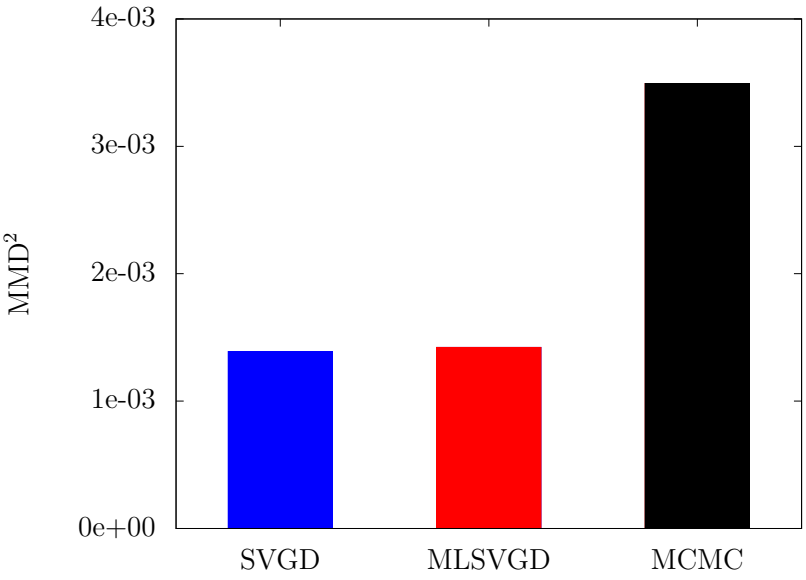


Figure 3.13: The estimated squared MMD using the estimator (3.82). The MLSVGD approximation has a comparable MMD to the original SVGD with the high-fidelity model only. Both have a lower MMD than MCMC suggesting higher quality samples.

Chapter 4

Conclusion and Outlook

4.1 Summary of contributions

To perform inference of an intractable target distribution that cannot be directly sampled, we may learn a tractable approximation to sample from instead. For multifidelity approaches to inference, the approximation may be derived using only a single surrogate model or through a hierarchy of increasingly accurate surrogate models as in multilevel methods. The quality of the learned approximation, often quantified through a probability divergence or metric, dictates the computational costs, or amount of sampling effort, that is required by the sampling procedure to achieve an estimate of the quantity of interest (2.1) within some error tolerance. Moreover, the measure of how good the learned approximation is depends on the fidelity of the surrogate model, leading to a trade-off between the costs of using a better surrogate model for training and the costs of evaluating the high-fidelity model to maintain accuracy of the outer-loop result (inferred quantity of interest).

In Chapter 2 we presented context-aware importance sampling, that selects an optimal, context-aware, surrogate model to minimize the costs of learning the biasing density and then

re-weighting samples according to the high-fidelity density. In this case the approximation took the form of the Laplace approximation to the context-aware surrogate density and the sampling effort required was determined by the χ^2 divergence to the target density. Both a theoretical analysis of the cost complexity and numerical results demonstrate that using the adaptive context-aware surrogate model can lead to runtime speedups of up to one order of magnitude over using a fixed surrogate model.

In Chapter 3 we considered a multilevel extension to SVGD that uses a hierarchy of surrogate models to build a sequence of increasingly accurate approximations as opposed to selecting a single optimal approximation. Here the sequence of approximations were given by integrating SVGD with respect to the surrogate distributions $\pi^{(1)}, \ldots, \pi^{(L-1)}$. The accuracy of these approximations were measured through the KL divergence and the sampling effort required at each level, including the final high-fidelity level, is determined by the KL divergence of the initial density to the target density. For MLSVGD, the surrogate models successfully learn a good approximation that serves as the initialization for the final level. This is again shown through both theoretical cost complexity bounds as well as several numerical examples, including a challenging 25-dimensional problem of inferring a glacier ice model, that exhibit up to an order of magnitude in speedup over traditional SVGD at the highest level.

4.2 Future work

Multifidelity inference is a rich area with many possible directions to either improve and analyze existing methods or to develop new ones.

More expressive approximations

The decomposition (1.5) of sampling effort in the example of Section 1.2.1 shows that increasing the capacity of the family of approximating densities can lead to better approxima-

tions. Two prominent examples of this are normalizing flows [Tabak and Turner, 2012, Tabak and Vanden-Eijnden, 2010, Gabrié et al., 2022] and transport maps [Moselhy and Marzouk, 2012, Parno and Marzouk, 2018]. Both of these parametric families of densities can provide more expressive approximations and can be trained efficiently. However, these methods lack approximation guarantees and finding the global optimum is challenging as their objective functions are non-convex. Despite this, both transport maps and normalizing flows have seen tremendous success in practice.

Extending existing methods

Another line of research to consider is extending the capabilities of existing methods. For example, in context-aware importance sampling we do not optimize over the number of offline evaluations N_0, but rather consider it to be a fixed constant. Translating approximation bounds in terms of N_0 would allow us to further exploit trade-off between online and offline costs. Alternatively, we only considered a single optimal surrogate model when perhaps it would be more advantageous to use multiple surrogate models as in sequential importance sampling [Liu, 2004, Latz et al., 2018]. For MLSVGD one could replace SVGD with one of its variants, for example the Stein variational Newton (SVN) method [Detommaso et al., 2018] or using a more adaptive kernels [Wang et al., 2019], in a straightforward manner. In this thesis we were primarily interested in inverse uncertainty quantification. However, one could also apply these methods to forward uncertainty propagation, particularly rare-event estimation. Such a transition would require careful adjustments of how the approximating densities are chosen.

Analyzing existing methods

Yet another potential direction to consider, related to the previous one discussed, involves building up analysis of existing methods. A more in-depth study of how to select the needed rates, which currently are determined during an expensive pilot study, could greatly improve the practicality of not only these methods, but multifidelity methods in general. For SVGD

there is still limited convergence analysis. Most of the convergence analysis is in the mean-field limit, a disconnect from practical implementations that use finite sample sizes. Such a convergence theory remains largely elusive for interacting particle methods and is a problem for the analysis of ensemble Kalman inversion [Iglesias et al., 2013, Schillings and Stuart, 2017] as well. As a separate example, there has been growing literature on the convergence guarantees of Langevin methods for log-concave sampling [Dwivedi et al., 2019, Brosse et al., 2017, Bubeck et al., 2015]. Understanding both the convergence of a sampling procedure as well as approximation guarantees can help us determine how to best allocate computational resources and minimize sampling effort.

APPENDIX A

Sub-Gaussian results

A.1 Orlicz norm of a Gaussian vector

Let $\boldsymbol{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{d\times d})$. Because, $\boldsymbol{X}$ is rotationally symmetric, for any unit vector $\boldsymbol{v}$, $\boldsymbol{v}^\top \boldsymbol{X} \sim N(0, \sigma^2)$. Thus, if $X \sim N(0, \sigma^2)$ we know that

$$\|\boldsymbol{X}\|_{\psi_2} = \|X\|_{\psi_2} .$$

By integrating directly, we have that

$$\begin{aligned}
\mathbb{E}\left[\exp(X^2/t^2)\right] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} \exp\left(-\frac{x^2}{2\sigma^2} + \frac{x^2}{t^2}\right) \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} \exp\left(-\frac{x^2}{2}\left(\frac{1}{\sigma^2} - \frac{2}{t^2}\right)\right) \mathrm{d}x \\
&= \frac{1}{\sigma\sqrt{\frac{1}{\sigma^2} - \frac{2}{t^2}}} \\
&= \frac{1}{\sqrt{1 - 2\sigma^2/t^2}} .
\end{aligned}$$

Setting the last line equal to 2 and solving for t gives the Orlicz norm

$$\|\boldsymbol{X}\|_{\psi_2} = \sqrt{\frac{8}{3}}\sigma\,.$$

A.2 Proof of Lemma 1

Proof. Suppose that $\boldsymbol{X}$ is a sub-Gaussian random vector and consider the matrix $\mathbf{A}$ to be a multiple of the identity, $\mathbf{A} = \alpha\mathbf{I}$ with $\alpha > 0$. We now only need to show that there exists an $\alpha > 0$ such that for all $\boldsymbol{m} \in \mathbb{R}^d$

$$\mathbb{E}_\eta\left[\exp\left(\alpha\|\boldsymbol{X}-\boldsymbol{m}\|^2\right)\right] = \mathbb{E}_\eta\left[\exp\left((\boldsymbol{X}-\boldsymbol{m})^T\mathbf{A}(\boldsymbol{X}-\boldsymbol{m})\right)\right] < \infty\,.$$

Since $\|\boldsymbol{v}+\boldsymbol{w}\|^2 \leq 2\|\boldsymbol{v}\|^2 + 2\|\boldsymbol{w}\|^2$ by the triangle inequality and the fact that $(a+b)^2 \leq 2a^2+2b^2$, we get the upper bound

$$\begin{aligned}\mathbb{E}_\eta\left[\exp\left(\alpha\|\boldsymbol{X}-\boldsymbol{m}\|^2\right)\right] &\leq \mathbb{E}_\eta\left[\exp\left(2\alpha\|\boldsymbol{m}\|^2 + 2\alpha\|\boldsymbol{X}\|^2\right)\right] \\ &= \exp\left(2\alpha\|\boldsymbol{m}\|^2\right)\mathbb{E}_\eta\left[\exp\left(2\alpha\|\boldsymbol{X}\|^2\right)\right]\,.\end{aligned}$$

Therefore, we now need to find $\alpha > 0$ such that

$$\mathbb{E}_\eta\left[\exp\left(2\alpha\|\boldsymbol{X}\|^2\right)\right] < \infty\,.$$

We now use the assumption that $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ is sub-Gaussian by taking the marginals

$$\begin{aligned}
\mathbb{E}_\eta \left[\exp\left(2\alpha \|\boldsymbol{X}\|^2\right) \right] &= \mathbb{E}_\eta \left[\exp\left(2\alpha \sum_{i=1}^d X_i^2 \right) \right] \\
&= \mathbb{E}_\eta \left[\exp\left(2\alpha \sum_{i=1}^d |\boldsymbol{e}_i^T \boldsymbol{X}|^2 \right) \right] \\
&= \mathbb{E}_\eta \left[\prod_{i=1}^d \exp\left(2\alpha |\boldsymbol{e}_i^T \boldsymbol{X}|^2\right) \right] ,
\end{aligned}$$

where $\boldsymbol{e}_i$ is the i-th canonical basis vector. We proceed by induction on the dimension d and make repeated use of the Cauchy-Schwarz inequality to show that this expectation is finite. When $d = 1$, take α_1 such that $\frac{1}{\sqrt{2\alpha_1}} > \|\boldsymbol{X}\|_{\psi_2}$ so that

$$\mathbb{E}_\eta \left[\exp\left(2\alpha_1 |\boldsymbol{e}_1^\top \boldsymbol{X}|^2\right) \right] = \mathbb{E}_\eta \left[\exp\left(\frac{|\boldsymbol{e}_1^\top \boldsymbol{X}|^2}{(1/\sqrt{2\alpha_1})^2} \right) \right] \leq 2 \, .$$

Note that since $\boldsymbol{X}$ is sub-Gaussian $\|\boldsymbol{X}\|_{\psi_2} < \infty$ we can indeed find an $\alpha_1 > 0$ to satisfy the inequality. Now suppose that for dimension $d-1$ there exists an α_{d-1} such that

$$\mathbb{E}_\eta \left[\prod_{i=1}^{d-1} \exp\left(2\alpha_{d-1} |\boldsymbol{e}_i^\top \boldsymbol{X}|^2\right) \right] = C_{d-1} < \infty \, .$$

By using the Cauchy-Schwarz inequality, we get that

$$\mathbb{E}_\eta \left[\prod_{i=1}^{d} \exp\left(2\alpha_{d} |\boldsymbol{e}_i^\top \boldsymbol{X}|^2\right) \right] \leq \mathbb{E}_\eta \left[\prod_{i=1}^{d-1} \exp\left(4\alpha_{d} |\boldsymbol{e}_i^\top \boldsymbol{X}|^2\right) \right]^{1/2} \mathbb{E}_\eta \left[\exp\left(4\alpha_{d} |\boldsymbol{e}_d^\top \boldsymbol{X}|^2\right) \right]^{1/2} .$$

Taking $\alpha_d \leq \alpha_{d-1}/2$ gives

$$\mathbb{E}_\eta \left[\prod_{i=1}^{d-1} \exp\left(4\alpha_{d} |\boldsymbol{e}_i^\top \boldsymbol{X}|^2\right) \right]^{1/2} \leq \mathbb{E}_\eta \left[\prod_{i=1}^{d-1} \exp\left(2\alpha_{d-1} |\boldsymbol{e}_i^\top \boldsymbol{X}|^2\right) \right]^{1/2} = C_{d-1}^{1/2} \, .$$

Taking α_d such that $\frac{1}{\sqrt{4\alpha_d}} > \|\boldsymbol{X}\|_{\psi_2}$ gives

$$\mathbb{E}_\eta \left[\exp\left(4\alpha_d |\boldsymbol{e}_d^\top \boldsymbol{X}|^2\right)\right]^{1/2} \leq \mathbb{E}_\eta \left[\exp\left(\frac{|\boldsymbol{e}_d^\top \boldsymbol{X}|^2}{(1/\sqrt{4\alpha_d})^2}\right)\right]^{1/2} \leq \sqrt{2}\,.$$

Thus, take $\alpha_d < \frac{1}{4}\min\{2\alpha_{d-1},\ \|\boldsymbol{X}\|_{\psi_2}^{-2}\}$, so that

$$\mathbb{E}_\eta \left[\prod_{i=1}^{d} \exp\left(2\alpha_d |\boldsymbol{e}_i^\top \boldsymbol{X}|^2\right)\right] \leq \sqrt{2C_{d-1}} < \infty\,.$$

Since the dimension is finite, we know that we will always be able to take $\alpha_d > 0$. Setting $\alpha = \alpha_d$, shows the first direction of the lemma.

For the converse suppose that there exists a symmetric positive-definite matrix $\mathbf{A} \succ 0$ so that for all vectors $\boldsymbol{m}$

$$\mathbb{E}_\eta \left[\exp\left((\boldsymbol{X}-\boldsymbol{m})^\top \mathbf{A}(\boldsymbol{X}-\boldsymbol{m})\right)\right] < \infty\,.$$

In particular, for $\boldsymbol{m} = \mathbf{0}$

$$\mathbb{E}_\eta \left[\exp\left(\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}\right)\right] = C < \infty\,.$$

For any $\boldsymbol{v} \in S^{d-1}$, we have that

$$\mathbb{E}_\eta \left[\exp\left(\frac{|\boldsymbol{v}^\top \boldsymbol{X}|^2}{t^2}\right)\right] \leq \mathbb{E}_\eta \left[\exp\left(\frac{\|\boldsymbol{X}\|^2}{t^2}\right)\right]\,,$$

since $|\boldsymbol{v}^\top \boldsymbol{X}| \leq \|\boldsymbol{v}\|\,\|\boldsymbol{X}\|$. Also, since the minimum eigenvalue satisfies $\lambda_{\min}^{\mathbf{A}} \leq \frac{\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}}{\|\boldsymbol{X}\|^2}$ for all $\boldsymbol{X} \neq \mathbf{0}$, we get

$$\mathbb{E}_\eta \left[\exp\left(\frac{\|\boldsymbol{X}\|^2}{t^2}\right)\right] \leq \mathbb{E}_\eta \left[\exp\left(\frac{\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}}{\lambda_{\min}^{\mathbf{A}} t^2}\right)\right] = \mathbb{E}_\eta \left[\left\{\exp\left(\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}\right)\right\}^{1/\lambda_{\min}(\mathbf{A})t^2}\right]\,.$$

If $\lambda_{\min}(\mathbf{A})t^2 > 1$, then the function

$$g(x) = x^{1/(\lambda_{\min}(\mathbf{A})t^2)}$$

is concave and increasing in x. By Jensen's inequality, we obtain

$$\mathbb{E}_\eta \left[\left\{ \exp\left(\boldsymbol{X}^\top \mathbf{A} \boldsymbol{X}\right) \right\}^{1/\lambda_{\min}^{\mathbf{A}} t^2} \right] \leq \mathbb{E}_\eta \left[\exp\left(\boldsymbol{X}^T \mathbf{A} \boldsymbol{X}\right) \right]^{1/\lambda_{\min}(\mathbf{A})t^2} = C^{1/\lambda_{\min}(\mathbf{A})t^2} .$$

Setting $C^{1/\lambda_{\min}(\mathbf{A})t^2} \leq 2$ and solving for t gives

$$t \geq \sqrt{\frac{\log C}{\lambda_{\min}(\mathbf{A}) \log 2}} .$$

Since this inequality holds for every $\boldsymbol{v} \in S^{d-1}$ we know that $\|\boldsymbol{X}\|_{\psi_2} < \infty$ and hence $\boldsymbol{X}$ is sub-Gaussian. □

APPENDIX B

Chi-squared divergence for Gaussian distributions

Let $\pi = N(\boldsymbol{m}, \boldsymbol{\Sigma})$ and $\mu = N(\boldsymbol{m}_0, \boldsymbol{\Sigma}_0)$. The chi-squared divergence is

$$\chi^2 (\pi \mid\mid \mu) = \int_{\mathbb{R}^d} \frac{\pi(\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})} \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} - 1 \,. \tag{B.1}$$

Plugging in the density functions gives

$$\int_{\mathbb{R}^d} \frac{|\boldsymbol{\Sigma}_0|^{1/2}}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|} \exp\left(-\frac{1}{2} Q(\boldsymbol{\theta})\right) \mathrm{d}\boldsymbol{\theta} - 1 \,. \tag{B.2}$$

where the quadratic exponent is

$$Q(\boldsymbol{\theta}) = 2(\boldsymbol{\theta} - \boldsymbol{m})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta} - \boldsymbol{m}) - (\boldsymbol{\theta} - \boldsymbol{m}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\theta} - \boldsymbol{m}_0) \,.$$

Combining powers of x within the exponent gives

$$Q(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top (2\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_0^{-1}) \boldsymbol{\theta} - 2 \left(2\boldsymbol{m}^\top \boldsymbol{\Sigma}^{-1} - \boldsymbol{m}_0^\top \boldsymbol{\Sigma}_0^{-1}\right) \boldsymbol{\theta} + \left(2\boldsymbol{m}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{m} - \boldsymbol{m}_0^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{m}_0\right) \,. \tag{B.3}$$

Now set $\mathbf{W} = 2\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_0^{-1}$,

$$\mathbf{w} = \mathbf{W}^{-1}\left(2\boldsymbol{\Sigma}^{-1}\boldsymbol{m} - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0\right),$$

and the constant

$$R = 2\boldsymbol{m}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{m} - \boldsymbol{m}_0^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0 - \mathbf{w}^\top\mathbf{W}\mathbf{w}\,,$$

so that the expression (B.3) becomes

$$Q(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \mathbf{w})^\top\mathbf{W}(\boldsymbol{\theta} - \mathbf{w}) + R\,.$$

Plugging into (B.2) gives

$$\begin{aligned}\chi^2\left(\pi \,||\, \mu\right) &= \int_{\mathbb{R}^d}\frac{\pi(\boldsymbol{\theta})}{\mu(\boldsymbol{\theta})}\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} - 1 \\ &= \frac{|\boldsymbol{\Sigma}_0|^{1/2}}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|}\exp\left(-\frac{1}{2}R\right)\int_{\mathbb{R}^d}\exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \mathbf{w})^\top\mathbf{W}(\boldsymbol{\theta} - \mathbf{w})\right)\mathrm{d}\boldsymbol{\theta} - 1\,,\end{aligned} \tag{B.4}$$

which is a Gaussian integral with mean $\mathbf{w}$ and covariance $\mathbf{W}^{-1}$ and becomes

$$\chi^2\left(\pi \,||\, \mu\right) = \frac{|\boldsymbol{\Sigma}_0|^{1/2}}{|\boldsymbol{\Sigma}| \cdot |2\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_0^{-1}|^{1/2}}\exp\left(-\frac{1}{2}R\right) - 1\,. \tag{B.5}$$

In the case where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$, the χ^2 divergence may then be simplified in this case

$$\begin{aligned}\chi^2\left(\pi \,||\, \mu\right) &= \exp\left(-\frac{1}{2}R\right) - 1 \\ &= \exp\left(\left(\boldsymbol{m} - \boldsymbol{m}_0\right)^\top\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{m} - \boldsymbol{m}_0\right)\right) - 1\,,\end{aligned} \tag{B.6}$$

which is zero exactly when $\boldsymbol{m} = \boldsymbol{m}_0$ and closely corresponds to the exponential bound.

Bibliography

[Agapiou et al., 2017] Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. (2017). Importance sampling: Intrinsic dimension and computational cost. *Statist. Sci.*, 32(3):405–431.

[Akyildiz and Míguez, 2021] Akyildiz, Ö. and Míguez, J. (2021). Convergence rates for optimised adaptive importance samplers. *Statistics and Computing*, 31(12).

[Al-Qaq et al., 1995] Al-Qaq, W. A., Devetsikiotis, M., and Townsend, J. K. (1995). Stochastic gradient optimization of importance sampling for the efficient simulation of digital communication systems. *IEEE Transactions on Communications*, 43(12):2975–2985.

[Alnaes et al., 2015] Alnaes, M. S., Blechta, J., Hake, J., Johansson, A., Kehlet, B., Logg, A., Richardson, C., Ring, J., Rognes, M. E., and Wells, G. N. (2015). The fenics project version 1.5.

[Alsup et al., 2022] Alsup, T., Hartland, T., Peherstorfer, B., and Petra, N. (2022). Further analysis of multilevel Stein variational gradient descent with an application to the Bayesian inference of glacier ice models. *pre-print*. arXiv:2212.03366.

[Alsup and Peherstorfer, 2022] Alsup, T. and Peherstorfer, B. (2022). Context-aware surrogate modeling for balancing approximation and sampling costs in multi-fidelity importance sampling and bayesian inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*.

[Alsup et al., 2021] Alsup, T., Venturi, L., and Peherestorfer, B. (2021). Multilevel Stein variational gradient descent with applications to Bayesian inverse problems. In Bruna, J., Hesthaven, J. S., and Zdeborova, L., editors, *Proceedings of Machine Learning Research*, volume 145 of *2nd Annual Conference on Mathematical and Scientific Machine Learning*, pages 1–25.

[Antoulas, 2005] Antoulas, A. (2005). *Approximation of Large-Scale Dynamical Systems*. Advances in Design and Control. SIAM.

[Antoulas et al., 2020] Antoulas, A. C., Beattie, C. A., and Gugercin, S. (2020). *Interpolatory Methods for Model Reduction*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

[Ba et al., 2019] Ba, J., Erdogdu, M., Ghassemi, M., Suzuki, T., and Wu, D. (2019). Towards characterizing the high-dimensional bias of kernel-based particle inference algorithms. In *2nd Symposium on Advances in Approximate Bayesian Inference*, pages 1–17.

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

[Benner et al., 2020] Benner, P., Goyal, P., Kramer, B., Peherstorfer, B., and Willcox, K. (2020). Operator inference for non-intrusive model reduction of systems with non-polynomial nonlinear terms. *Computer Methods in Applied Mechanics and Engineering*, 372.

[Benner et al., 2015] Benner, P., Gugercin, S., and Willcox, K. (2015). A survery of projection-based model reduction for parametric dynamical systems. *SIAM Rev.*, 57(4):483–531.

[Beskos et al., 2017] Beskos, A., Jasra, A., Law, K., Tempone, R., and Zhou, Y. (2017). Multilevel sequential Monte Carlo samplers. *Stochastic Processes and their Applications*, 127(5):1417 – 1440.

[Brenner and Scott, 2008] Brenner, S. and Scott, R. (2008). *The Mathematical Theory of Finite Element Methods*. Texts in Applied Mathematics. Springer-Verlag New York.

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

[Brosse et al., 2017] Brosse, N., Durmus, A., Moulines, E., and Pereyra, M. (2017). Sampling from a log-concave distribution with compact support with proximal langevin monte carlo. In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 319–342. PMLR.

[Bruna et al., 2022] Bruna, J., Peherstorfer, B., and Vanden-Eijnden, E. (2022). Neural galerkin scheme with active learning for high-dimensional evolution equations. *arXiv:2203.01360*.

[Bubeck et al., 2015] Bubeck, S., Eldan, R., and Lehec, J. (2015). Finite-time analysis of projected langevin monte carlo. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

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

[Cao et al., 2011] Cao, Y., Gunzburger, M., Hua, F., and Wang, X. (2011). Analysis and finite element approximation of a coupled, continuum pipe-flow/Darcy model for flow in porous media with embedded conduits. *Numerical Methods Partial Differential Equations*, 27:1242–1252.

[Chatterjee and Diaconis, 2018] Chatterjee, S. and Diaconis, P. (2018). The sample size required in importance sampling. *Ann. Appl. Probab.*, 28(2):1099–1135.

[Chen and Quarteroni, 2013] Chen, P. and Quarteroni, A. (2013). Accurate and efficient evaluation of failure probability for partial differential equations with random input data. *Computer Methods in Applied Mechanics and Engineering*, 267:233–260.

[Chen et al., 2017] Chen, P., Quarteroni, A., and Rozza, G. (2017). Reduced basis methods for uncertainty quantification. *SIAM/ASA J. Uncertain. Quantif.*, 5:813–869.

[Chen et al., 2019] Chen, P., Wu, K., Chen, J., Leary-Roseberry, T. O., and Ghattas, O. (2019). Projected Stein variational Newton: A fast and scalable Bayesian inference method in high dimensions. In Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 15130–15139. Curran Associates, Inc.

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

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

[Cliffe et al., 2011] Cliffe, K. A., Giles, M. B., Scheichl, R., and Teckentrup, A. L. (2011). Multilevel monte carlo methods and applications to elliptic pdes with random coefficients. *Computing and Visualization in Science*, 14(1):3.

[Cotter et al., 2013] Cotter, S. L., Roberts, G. O., Stuart, A. M., and White, D. (2013). MCMC methods for functions: modifying old algorithms to make them faster. *Statist. Sci.*, 28(3):424–446.

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

[Dissanayake and Phan-Thien, 1994] Dissanayake, M. W. M. G. and Phan-Thien, N. (1994). Neural-network-based approximations for solving partial differential equations. *Communications in Numerical Methods in Engineering*, 10(3):195–201.

[Dodwell et al., 2015] Dodwell, T. J., Ketelsen, C., Scheichl, R., and Teckentrup, A. L. (2015). 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.

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

[Dwivedi et al., 2019] Dwivedi, R., Chen, Y., Wainwright, M. J., and Yu, B. (2019). Log-concave sampling: Metropolis-hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183):1–42.

[Farcas et al., 2022] Farcas, I., Peherstorfer, B., Neckel, T., Jenko, F., and Bungartz, H.-J. (2022). Context-aware learning of hierarchies of low-fidelity models for multi-fidelity uncertainty quantification. *pre-print*. arXiv:2211.10835.

[Farcas, 2020] Farcas, I. G. (2020). *Context-aware model hierarchies for higher-dimensional uncertainty quantification*. PhD thesis, Technische Universität München.

[Forrester and Keane, 2009] Forrester, A. and Keane, A. (2009). Recent advances in surrogate-based optimization. *Progr. Aerosp. Sci.*, 45:50–79.

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

[Gabrié et al., 2022] Gabrié, M., Rotskoff, G., and Vanden-Eijnden, E. (2022). Adaptive Monte Carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences*, 119.

[Gallego and Insua, 2020] Gallego, V. and Insua, D. (2020). Stochastic gradient mcmc with repulsive forces. *arXiv:1812.00071*.

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

[Goodman and Weare, 2010] Goodman, J. and Weare, J. (2010). Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science*.

[Gorodetsky et al., 2020a] Gorodetsky, A., Geraci, G., Eldred, M., and Jakeman, J. (2020a). A generalized approximate control variate framework for multifidelity uncertainty quantification. *Journal of Computational Physics*, 408.

[Gorodetsky et al., 2020b] Gorodetsky, A., Jakeman, J., Geraci, G., , and Eldred, M. (2020b). MFNets: Multi-fidelity data-driven networks for Bayesian learning and prediction. Technical report, Sandia National Laboratories.

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

[Gretton et al., 2012] Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773.

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

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

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

[Haji-Ali et al., 2016] Haji-Ali, A., Nobile, F., and Tempone, R. (2016). Multi-index Monte Carlo: when sparsity meets sampling. *Numer. Math.*, 132(4):767–806.

[Han and Liu, 2018] Han, J. and Liu, Q. (2018). Stein variational gradient descent without gradient. In *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.*, volume 80.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

[Hesthaven et al., 2016] Hesthaven, J. S., Rozza, G., and Stamm, B. (2016). *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*. SpringerBriefs in Mathmatics. Springer International Publishing.

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

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

[Iglesias et al., 2013] Iglesias, M. A., Law, K. J. H., and Stuart, A. M. (2013). Ensemble Kalman methods for inverse problems. *Inverse Problems*, 29(4).

[Ionita and Antoulas, 2014] Ionita, A. C. and Antoulas, A. C. (2014). Data-driven parametrized model reduction in the Loewner framework. *SIAM Journal on Scientific Computing*, 36(3):A984–A1007.

[Jasra et al., 2017] Jasra, A., Kamatani, K., Law, K., and Zhou, Y. (2017). Multilevel particle filters. *SIAM Journal on Numerical Analysis*, 55(6):3068–3096.

[Jordan et al., 1998] Jordan, R., Kinderlehrer, D., and Otto, F. (1998). The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1).

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

[Konrad et al., 2021] Konrad, J., Farcas, I., Peherstorfer, B., Siena, A., Jenko, F., Neckel, T., and Bungartz, H. (2021). Data-driven low-fidelity models for multi-fidelity monte carlo sampling in plasma micro-turbulence analysis. *Journal of Computational Physics*.

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

[Kramer et al., 2019] Kramer, B., Marques, A., Peherstorfer, B., Villa, U., and Willcox, K. (2019). Multifidelity probability estimation via fusion of estimators. *Journal of Computational Physics*, 392:385–402.

[Lam et al., 2015] Lam, R., Allaire, D., and Willcox, K. (2015). Multifidelity optimization using statistical surrogate modeling for non-hierarchical information sources. United States. Air Force. Office of Scientific Research. Multidisciplinary University Research Initiative (Grant FA9550- 09-0613).

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

[Leimkuhler et al., 2018] Leimkuhler, B., Matthews, C., and Weare, J. (2018). Ensemble preconditioning for Markov chain Monte Carlo simulation. *Statistics and Computing*, 28(2):277–290.

[LeVeque, 2007] LeVeque, R. J. (2007). *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*. SIAM.

[Leviyev et al., 2022] Leviyev, A., Chen, J., Wang, Y., Ghattas, O., and Zimmerman, A. (2022). A stochastic Stein variational Newton method. *arXiv:2204.09039*.

[Li et al., 2011] Li, J., Li, J., and Xiu, D. (2011). An efficient surrogate-based method for computing rare failure probability. *Journal of Computational Physics*, 230(24):8683–8697.

[Li and Xiu, 2010] Li, J. and Xiu, D. (2010). Evaluation of failure probability via surrogate models. *Journal of Computational Physics*, 229(3):8966–8980.

[Li et al., 2021] Li, Z., Fan, Y., and Ying, L. (2021). Multilevel fine-tuning: Closing generalization gaps in approximation of solution maps under a limited budget for training data. *Multiscale Modeling & Simulation*, 19(1).

[Liu et al., 2019] Liu, C., Zhuo, J., Cheng, P., Zhang, R., and Zhu, J. (2019). Understanding and accelerating particle-based variational inference. In Chaudhuri, K. and Salakhutdinov, R., 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. PMLR.

[Liu, 2004] Liu, J. S. (2004). *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer New York, NY, 1 edition.

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

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

[Lu et al., 2019] Lu, J., Lu, Y., and Nolen, J. (2019). Scaling limit of the Stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671.

[Ma et al., 2015] Ma, Y., Chen, T., and Fox, E. (2015). A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pages 2917–2925.

[Majda and Gershgorin, 2010] Majda, A. J. and Gershgorin, G. (2010). Quantifying uncertainty in climate change science through empirical information theory. *Proceedings of the National Academy of Sciences of the United States of America*, 107(34):14958–14963.

[Marzouk et al., 2016] Marzouk, Y., Moselhy, T., Parno, M., and Spantini, A. (2016). Sampling via measure transport: An introduction.

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

[Maurais, 2022] Maurais, A. (2022). Multifidelity covariance estimation three ways. Master's thesis, Massachusetts Institute of Technology.

[Morokoff and Caflisch, 1995] Morokoff, W. J. and Caflisch, R. E. (1995). Quasi-Monte Carlo integration. *J. Comput. Phys.*, 122(2):218–230.

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

[Ng and Willcox, 2016] Ng, L. W. and Willcox, K. (2016). Monte-Carlo information-reuse approach to aircraft conceptual design optimization under uncertainty. *Journal of Aircraft*, 53:427–438.

[Nobile et al., 2008] Nobile, F., Tempone, R., and Webster, C. (2008). A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 46(5):2309–2345.

[Nocedal and Wright, 2006] Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, NY, 2 edition.

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

[Pattyn et al., 2008] Pattyn, F., Perichon, L., Aschwanden, A., Breuer, B., de Smedt, B., Gagliardini, O., Gudmundsson, G. H., Hindmarsh, R. C. A., Hubbard, A., Johnson, J. V., Kleiner, T., Konovalov, Y., Martin, C., Payne, A. J., Pollard, D., Price, S., Ruckamp, M., Saito, F., Soucek, O., Sugiyama, S., , and Zwinger, T. (2008). Benchmark experiments for higher-order and full-Stokes ice sheet models (ISMIP-HOM). *The Cryosphere*, 2:95 – 108.

[Peherstorfer, 2019] Peherstorfer, B. (2019). Multifidelity Monte Carlo estimation with adaptive low-fidelity models. *SIAM/ASA Journal on Uncertainty Quantification*, 7:579–603.

[Peherstorfer et al., 2018a] Peherstorfer, B., Beran, P., and Willcox, K. (2018a). Multifidelity monte carlo estimation for large-scale uncertainty propagation. In *2018 AIAA Non-Deterministic Approaches Conference*. AIAA.

[Peherstorfer et al., 2016a] Peherstorfer, B., Cui, T., Marzouk, Y., and Willcox, K. (2016a). Multifidelity importance sampling. *Computer Methods in Applied Mechanics and Engineering*, 300:490–509.

[Peherstorfer et al., 2018b] Peherstorfer, B., Gunzburger, M., and Willcox, K. (2018b). Convergence analysis of multifidelity monte carlo estimation. *Numerische Mathematik*, 139(3):683–707.

[Peherstorfer et al., 2017] Peherstorfer, B., Kramer, B., and Willcox, K. (2017). Combining multiple surrogate models to accelerate failure probability estimation with expensive high-fidelity models. *Journal of Computational Physics*, 341:61–75.

[Peherstorfer et al., 2018c] Peherstorfer, B., Kramer, B., and Willcox, K. (2018c). 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.

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

[Peherstorfer and Willcox, 2016] Peherstorfer, B. and Willcox, K. (2016). Data-driven operator inference for nonintrusive projection-based model reduction. *Computer Methods in Applied Mechanics and Engineering*, 306:196–215.

[Peherstorfer et al., 2016b] Peherstorfer, B., Willcox, K., and Gunzburger, M. (2016b). Optimal model management for multifidelity monte carlo estimation. *SIAM Journal on Scientific Computing*, 38(5):A3163–A3194.

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

[Petra et al., 2014] Petra, N., Martin, J., Stadler, G., and Ghattas, O. (2014). A computational framework for infinite-dimensional Bayesian inverse problems, part II: Stochastic Newton MCMC with application to ice sheet flow inverse problems. *SIAM J. Sci. Comput.*, 36(4):A1525–A1555.

[Qian et al., 2020] Qian, E., Kramer, B., Peherstorfer, B., and Willcox, K. (2020). Lift & learn: Physics-informed machine learning for large-scale nonlinear dynamical systems. *Physica D: Nonlinear Phenomena*, Volume 406.

[Quarteroni et al., 2011] Quarteroni, A., Rozza, G., and Manzoni, A. (2011). Certified reduced basis approximation for parametrized partial differential equations and applications. *Journal of Mathematics in Industry*, 1(1):1–49.

[Raissi et al., 2019] Raissi, M., Perdikaris, P., and Karniadakis, G. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707.

[Ranganath et al., 2014] Ranganath, R., Gerrish, S., and Blei, D. (2014). Black Box Variational Inference. In Kaski, S. and Corander, J., 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.

[Rasmussen and Williams, 2016] Rasmussen, C. and Williams, C. (2016). *Gaussian Processes for Machine Learning*. MIT Press.

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

[Robert and Casella, 2004] Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.

[Robinson et al., 2006] Robinson, T., Eldred, M., Willcox, K., and Haimes, R. (2006). Strategies for multifidelity optimization with variable dimensional hierarchical models. In *47th AIAA/ASME/ASCE*.

[Ryu and Boyd, 2014] Ryu, E. and Boyd, S. (2014). Adaptive importance sampling via stochastic convex programming. *pre-print*. arXiv:1412.4845.

[Sanz-Alonso, 2018] Sanz-Alonso, D. (2018). Importance sampling and necessary sample size: An information theory approach. *SIAM/ASA J. Uncertainty Quantification*, 6(2):867–879.

[Schaden and Ullmann, 2020] Schaden, D. and Ullmann, E. (2020). On multilevel best linear unbiased estimators. *SIAM/ASA J. Uncertainty Quantification*, 8(2):601–635.

[Schillings et al., 2020] Schillings, C., Sprungk, B., and Wacker, P. (2020). On the convergence of the Laplace approximation and noise-level-robustness of Laplace-based Monte Carlo methods for Bayesian inverse problems. *Numer. Math.*, 145:915–971.

[Schillings and Stuart, 2017] Schillings, C. and Stuart, A. M. (2017). Analysis of the ensemble Kalman filter for inverse problems. *SIAM J. Numer. Anal.*, 55(3):1264–1290.

[Shyamkumar et al., 2022] Shyamkumar, N., Gugercin, S., and Peherstorfer, B. (2022). Towards context-aware learning for control: Balancing stability and model-learning error. In *IEEE American Control Conference*.

[Sirignano and Spiliopoulos, 2018] Sirignano, J. and Spiliopoulos, K. (2018). DGM: A deep learning algorithm for solving partial differential equations. *J. Comput. Phys.*, 375:1339–1364.

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

[Sullivan, 2015] Sullivan, T. (2015). *Introduction to Uncertainty Quantification*. Springer.

[Swischuk et al., 2019] Swischuk, R., Mainini, L., Peherstorfer, B., and Willcox, K. (2019). Projection-based model reduction: Formulations for physics-based machine learning. *Computers & Fluids*, 179:704–717.

[Tabak and Turner, 2012] Tabak, E. and Turner, C. (2012). A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164.

[Tabak and Vanden-Eijnden, 2010] Tabak, E. and Vanden-Eijnden, E. (2010). Density estimation by dual ascent of the log-likelihood. *Commun. Math Sci.*, 8(1):217–233.

[Teckentrup et al., 2013] Teckentrup, A., Scheichl, R., Giles, M., and Ullmann, E. (2013). Further analysis of multilevel Monte Carlo methods for elliptic PDEs with random coefficients. *Numer. Math.*, 125:569–600.

[Trottenberg et al., 2001] Trottenberg, U., Oosterlee, C. W., Schuller, A., and Brandt, A. (2001). *Multigrid*. Academic Press.

[Tsybakov, 2009] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer.

[van der Vaart, 1998] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

[Vershynin, 2018] Vershynin, R. (2018). *High-dimensional probability: an introduction with applications in data science*. Cambridge University Press.

[Villa et al., 2016] Villa, U., Petra, N., and Ghattas, O. (2016). Documentation to "hIPPYlib: an Extensible Software Framework for Large-scale Deterministic and Bayesian Inverse Problems". http://hippylib.github.io.

[Villa et al., 2018] Villa, U., Petra, N., and Ghattas, O. (2018). hIPPYlib: an Extensible Software Framework for Large-scale Deterministic and Bayesian Inverse Problems. *Journal of Open Source Software*, 3(30).

[Villa et al., 2021] Villa, U., Petra, N., and Ghattas, O. (2021). HIPPYlib: An Extensible Software Framework for Large-Scale Inverse Problems Governed by PDEs: Part I: Deterministic Inversion and Linearized Bayesian Inference. *ACM Trans. Math. Softw.*, 47(2).

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

[Wang et al., 2019] Wang, D., Tang, Z., Bajaj, C., and Liu, Q. (2019). Stein variational gradient descent with matrix-valued kernels. In *Advances in Neural Information Processing Systems*, volume 32.

[Weissmann et al., 2022] Weissmann, S., Wilson, A., and Zech, J. (2022). Multilevel optimization for inverse problems. In *Proceedings of Machine Learning Research, Volume 178: Conference on Learning Theory, 2022*.

[Werner and Peherstorfer, 2022] Werner, S. and Peherstorfer, B. (2022). Context-aware controller inference for stabilizing dynamical systems from scarce data. *arXiv:2207.11049*.

[Zhang et al., 2019] Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. (2019). Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026.