GENERALISED VARIATIONAL INFERENCE IN INFINITE DIMENSIONS

Veit D. Wild
Keble College
University of Oxford

A thesis presented for the degree of
Doctor of Philosophy

Hillary 2024

Für meine Eltern Ilona und Matthias. Für mein Schwester Pia.

Acknowledgements

A doctorate is a long and challenging enterprise which is neither possible nor enjoyable, without the support of mentors, friends and family.

First and foremost I want to thank my supervisors, Dino Sejdinovic and George Deligiannidis. Dino, thank you for allowing me to freely explore my own research interest whilst always providing support and guidance when I needed it. You truly are a great scholar representing the best of academia for me. George, thank you for the support you have given at the end of my dissertation, when you became my second supervisor. I have always looked with great admiration at your research. A special thanks to Jeremias, my third supervisor in all but title and who worked relentlessly on improving my writing and making my ideas shine. I am very grateful for the support I received at UCL and enjoyed our discussion a lot!

I furthermore want to thank all my collaborators. My collaboration with George was one of the highlights, during my doctorate and truly showed me how enjoyable research can be. My work with Robert has shown me how fast one can make progress when two different strengths are combined. The paper with Sahra was wonderful, not only because of her amazing skills, but also because of the enjoyable conversations in our office. A special thanks to James, who has worked tirelessly in the past few weeks to run experiments.

I have also had the great fortune of being blessed with countless amazing friends. First, I want to thank Ben for hosting me at his place in London for weeks without ever uttering a single complaint. I am truly grateful for your friendship and thank you for the wonderful time we have together. I also want to thank the four people representing the center of my social life in Oxford: Alejandro, Alex, Michael and Santiago. Spending the weekends with you guys was always amazing and something I dearly miss now. A special thanks to my friends in Germany: Eva, Lukas, Robin and Sergej. You guys know how much I love you and enjoy spending time with you! Thanks also to Carlo, who played a crucial role in convincing me to study mathematics and has been an great friend ever since.

Last but not least I want to thank my family members. My parents Ilona and Matthias who have raised me in a house full of love, support and the freedom to pursue my own passions without any expectations. My sister Pia, who is one of my favourite people on this planet and a blessing to this world. My in-laws Rekha and Rajeev who have treated my like their own son from day one and gone above and beyond to support me.

And of course the deepest and most special thanks to my remarkable wife Kiran! She endured the three years of my doctoral studies whilst having received her doctorate already four years ago. Without your support I would not have been able to finish so quickly.

Declaration

I hereby declare that except where specific reference is made to the work of others, the intellectual contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification. The particulars of my personal contributions can be found in Section 1.2 of the introduction. This dissertation is my own work except as specified in the text.

Veit D. Wild

Hillary 2024

Abstract

This thesis develops a rigorous mathematical theory for variational inference (VI) and generalised variational inference (GVI) when the underlying parameter space is an infinite-dimensional function space. The insights gained from our improved theoretical understanding of VI and GVI allow us to propose novel approaches for solving the VI and GVI optimisation problem via introduction of infinite-dimensional variational parameters or infinite-dimensional gradient descent methods.

Although GVI and in particular VI have previously been employed for function space inference, the infinite-dimensional nature of the parameter space has so far been ignored. This has hindered the development of new inference approaches for which a deep understanding of the mathematical structure underpinning VI and GVI is a prerequisite.

This thesis closes this gap by advancing mathematical concepts from infinite-dimensional analysis and probability theory. In particular we use Gaussian random elements in Banach spaces to formalise VI and GVI for function space inference. Consequently, we can access new analytical tools—such as the Wasserstein gradient flow or parameterisations in the space of Gaussian measures—to solve the VI and GVI optimisation problem.

The result is a rigorous theory for GVI in function space and a plethora of competitive novel algorithms for uncertainty quantification such as Gaussian Wasserstein inference, deep repulsive Langevin ensembles and projected Langevin sampling. All methods are rigorously derived from the same GVI objective and our theory unifies several approaches for uncertainty quantification in function space as well as parameter space.

Table of Contents

1 | Introduction

1.1 Motivation

In the past decade, the field of machine learning has advanced with breathtaking speed, marked by monumental breakthroughs across diverse domains, including reinforcement learning, language modeling, and computer vision [Zhang and Lu, 2021]. Prominent models that have found widespread recognition in the public are numerous and include names like Alexnet [Krizhevsky et al., 2012], AlphaGo [Silver et al., 2016], AlphaFold [Jumper et al., 2021], StableDiffusion [Rombach et al., 2022], and GPT-4 [Achiam et al., 2023], among others.

This remarkable progress owes much of its success to the enhanced capacity to process increasingly vast and intricate datasets, coupled with more refined functional approximations. However, despite these impressive achievements, machine learning models still grapple with a critical limitation—they lack the ability to *know what they don't know*.

Instances of models exhibiting this limitation are as prevalent as the algorithms themselves. For instance, Alexnet can be deceived into misclassifying images through imperceptible alterations to human observers [Szegedy et al., 2013]. Similarly, ChatGPT, while proficient in generating coherent text, confidently generates fictitious legal cases to bolster its argumentation [Bohanon, 2023].

The field of uncertainty quantification is concerned with mitigating such shortcomings. Rather than solely focusing on (point) predictions, the goal is to empower machine learning models with a means of gauging uncertainty, such as a confidence score. This score could enable human users to distinguish between reliable and unreliable machine learning predictions and aid subsequent decision making.

The landscape of approaches aimed at providing uncertainty quantification is vast and complex, making it challenging to give a unifying account. Nevertheless, a recurring thread present in many of these approaches is as follows: when confronted with multiple plausible explanations for the observed data, significant divergence among various models in their predictions for unobserved data should serve as a robust indicator of high uncertainty. Conversely, when several distinct models consistently provide the same prediction for a novel, unobserved data point, a greater degree of confidence is warranted in those predictions.

Generalised variational inference (GVI) [Knoblauch, 2021] is a unifying approach for uncertainty quantification, offering a structured mathematical framework to formalise the heuristic described above. In GVI, the conventional loss-minimisation process of statistical machine learning is "lifted" into the space of

probability measures. This reformulation enables us to replace singular point predictions with probability measures over a range of possibilities. As a result, a measure of dispersion, such as variance, can readily be harnessed as a confidence score, allowing for the assessment of the prediction's reliability.

My contributions to the field of (generalised) variational inference are reflected by the two distinct meanings of the word 'infinite' in the title of this dissertation. Firstly, I develop a rigorous theory and proposed several algorithms for (generalised) variational inference where the underlying parameter space is an infinite-dimensional function space. Secondly, I pioneer the use of the Wasserstein gradient flow—a mathematical structure that can be conceptualised as a gradient flow in the infinite-dimensional space of [probability] measures—in the context of GVI on finite-dimensional parameter spaces and in the context of VI on infinite-dimensional parameter spaces.

The thesis therefore has two distinct but closely related goals. On one hand, we want to develop a rigorous theory for (generalised) variational inference on infinite-dimensional function spaces. On the other hand, we want to propose new inference algorithm that provide better solutions to the optimisation problem that underpins generalised variational inference both in finite and infinite dimensions. In this thesis, we will frequently encounter that these two goals are in perfect harmony and that a well-developed mathematical theory of a problem is often a stepping stone for novel and improved inference algorithms.

1.2 Thesis Layout and Statements of Authorship

The following outlines the content of each thesis chapter and provides a summary of the contributions from each author.

The chapter "Variational Gaussian Processes: A Functional Analysis View" introduces Gaussian random element regression and variational inference for Gaussian measures on Banach spaces. This work is intricately connected to Sparse Variational Gaussian Processes (SVGP) and offers fresh insights into various inducing feature approaches. Co-authored with George Wynne, this joint first-author paper resulted from a collaborative effort, with contributions from both authors seamlessly interwoven. The paper was presented at the Conference for Artificial Intelligence and Statistics (AISTATS) in 2022.

The chapter "Generalised Variational Inference in Function Space: Gaussian Measures Meet Bayesian Deep Learning" extends the GVI framework to infinite-dimensional Hilbert spaces, paving the way for a function space inference method based on regularisation with the Wasserstein distance. The resulting Gaussian Wasserstein inference method, a joint first-author paper with Robert Hu, and supervised by Dino Sejdinovic, demonstrates impressive performance on benchmark datasets. The theoretical work and proofs are my own contribution, while Robert conducted the experiments and developed the software

library. The paper was accepted at the Conference on Neural Information Processing Systems (NeurIPS) in 2022.

The chapter "A Rigorous Link Between Deep Ensembles and Variational Bayesian Methods" scrutinises GVI on the finite-dimensional Euclidean parameter space, implementing gradient descent directly in the infinite-dimensional space of probability measures. This approach enables the derivation and theoretical study of new sampling schemes under different regularisers, such as maximum mean discrepancy (MMD) and Kullback-Leibler divergence. Supervised by Dino Sejdinovic and Jeremias Knoblauch, with Sahra Ghalebikesabi leading the experiments and software library development, the paper was accepted for an oral presentation at the Conference on Neural Information Processing Systems (NeurIPS) in 2023.

The chapter "Bayesian Inference in Function Space via the Wasserstein Gradient Flow" develops a sampling procedure for Gaussian random elements by implementing the Wasserstein gradient flow for an infinite-dimensional Hilbert space. The latter can be simulated with the Langevin Stochastic differential equation in Hilbert space. We propose an approximate sampling technique that uses an orthogonal projection onto the first M-components of the spectral basis of the covariance operator. Interestingly, the algorithm as a special case recovers the posterior arising from the sparse variational Gaussian process (SVGP) [Titsias, 2009a]–owed to the fact that the same sufficiency assumption underlies both methods. However, whereas the SVGP posterior is parametrically constrained to be a Gaussian process, our method is based on a variational family that can freely explore the space of [(sufficiently regular)] probability measures on $\mathbb{R}^M$. Our method is provably close to the optimal M-dimensional variational approximation of the Bayesian posterior for convex and Lipschitz continuous negative log likelihoods. All theoretical contributions are my own. It was supervised by Dino Sejdinovic and Jeremias Knoblauch. James Wu was responsible for implementing the experiments. This is unpublished work, written in manuscript style.

2 | Generalised Variational Inference in a Nutshell

This chapter introduces the GVI optimisation problem which is the central object of study in my dissertation. I elaborate on the view of GVI as mathematical lifting of a finite-dimensional optimisation problem onto the space of probability measures.

I further discuss my contributions to the field of GVI and embedd them into the context of existing literature. This chapter provides a concise summary of my perspective on GVI and the motivations behind my work.

2.1 The Optimisation-Centric View on Generalised Variational Inference

The central object of this thesis is the minimisation problem for the functional $L : \mathcal{P}(\Theta) \to [-\infty, \infty]$ defined as

$$L(Q) := \int \ell(\theta)\, dQ(\theta) + \lambda D(Q, P). \tag{2.1}$$

Here, we denote the parameter space as Θ and the set of all probability measures on Θ as $\mathcal{P}(\Theta)$. In the above $\ell : \Theta \to \mathbb{R}$ is a *loss* function, $P \in \mathcal{P}(\Theta)$ a *prior or reference measure*, $D : \mathcal{P}(\Theta) \times \mathcal{P}(\Theta) \to [-\infty, \infty]$ a *discrepancy measure or regulariser* and $\lambda > 0$ a *scaling constant*.

We refer to the minimisation problem for L in (2.1) as GVI-problem. It can be guaranteed— under mild assumption on ℓ and D—that a unique global minimiser

$$Q^* := \arg\min \left\{ L(Q) \,:\, Q \in \mathcal{P}(\Theta) \right\} \tag{2.2}$$

for L exists (cf. Appendix C.1).

The formulation of the GVI-problem, as presented above, was introduced by Knoblauch [2021] as a comprehensive framework summarissing and extending various approaches to generalise (variational) Bayesian inference [Grünwald, 2011, Bissiri et al., 2016, Holmes and Walker, 2017, Grünwald and Van Ommen, 2017, Miller and Dunson, 2018, Nakagawa and Hashimoto, 2020, Chérief-Abdellatif and Alquier, 2020]. Specifically, for a Bayesian model with a prior P and likelihood $p(y|\theta)$, the Bayesian posterior can be identified as the solution to the GVI-problem Q^* when $\ell(\theta) = -\log p(y|\theta)$, D is the Kullback-Leibler divergence (KL) and $\lambda = 1$. Drawing from its roots in Bayesian inference, it is argued [Bissiri et al., 2016, Knoblauch, 2021] that, even in scenarios where $\ell(\theta) \neq -\log p(y|\theta)$ or $D \neq \mathrm{KL}$, solving the GVI problem encapsulates the process of updating prior beliefs about the parameter $\theta \in \Theta$, as

represented by P, to posterior beliefs represented by Q^*.

I have come to prefer a purely optimisation focused justification of GVI that avoids the rationale of updating prior to posterior beliefs. In my view, the GVI-problem in (2.1) is interpreted as being derived from a minimisation problem for $\ell : \Theta \to \mathbb{R}$ in two steps: probabilistic lifting and convexification (cf. Figure (2.1)). The first step can be seen as relaxation of the problem for ℓ which is rooted in the knowledge,

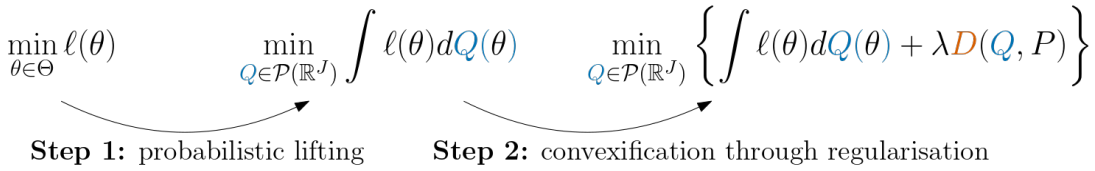


Figure 2.1: Illustration of convexification through probabilistic lifting.

that there is typically more than one parameter setting with low loss for ℓ to be found. The second step introduces a convex regulariser which singles out a unique probability measure amongst all candidate measures that lead to a low loss on average.

Hence, the global minimiser Q^* is interpreted as containing all the many minimisers of ℓ. This is visually illustrated in Figure 2.2, where the relationship between Q^* and the original loss ℓ is depicted. The loss

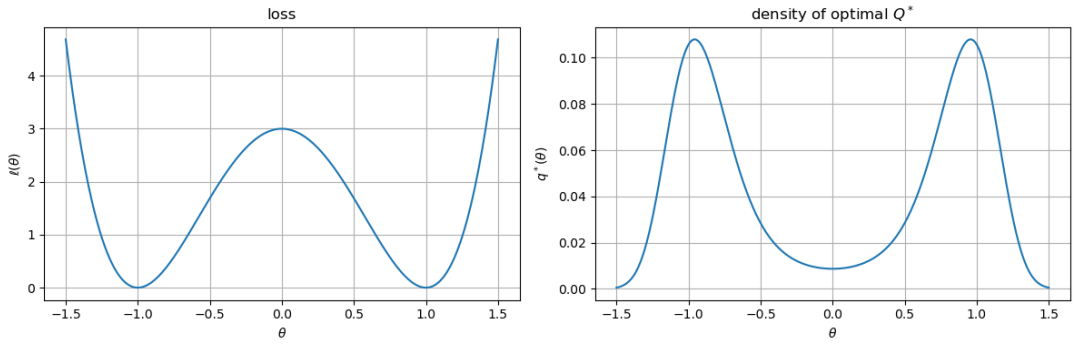


Figure 2.2: The figure illustrates how Q^* summarises the low-loss regions of ℓ. The loss is given as $\ell(\theta) = 3(\theta + 1)^2(\theta - 1)^2$, the regulariser is the Kullback-Leibler divergence, P is the standard normal distribution and $\lambda = 1$.

ℓ has two global minimisers and it is precisely in these regions where Q^* concentrates. The reference measure P can be used to encode knowledge about regions in the parameter space Θ that are more plausible. The scaling factor λ allows us to control the strength of regularisation.

This purely optimisation-centric view does not inherently provide us with an interpretation of the probabilities over parameter regions encoded in Q^* as posterior beliefs or updated beliefs. In my experience, the usage of such terminology leads to more confusion than necessary and is by no means required to justify an interest in solving the GVI-problem (2.1). In my view we simply query Q^* to generate points of

$\Theta^* := \arg\min\{\ell(\theta) : \theta \in \Theta\}$ and its vicinity. Finding Q^* is therefore a means to obtain access to the many minimisers of ℓ.

Moreover, the use of Q^* in the context of uncertainty quantification remains valid albeit with a new justification. To illustrate this consider the setting of statistical machine learning with squared loss ℓ given as

$$\ell(\theta) := \sum_{n=1}^{N} (y_n - f_\theta(x_n))^2 \tag{2.3}$$

for a set of input-output pairs $\mathcal{D} := \{(x_n, y_n) : n = 1, \ldots, N\} \subset \mathcal{X} \times \mathbb{R}$. Here $f_\theta : \mathcal{X} \to \mathbb{R}$ parameterises a set of candidate functions and a low loss is associated with a good fit for the points in $\mathcal{D}$. Given a new point $x \in \mathcal{X}$, we can leverage Q^* to induce a distribution over predictive outcomes[1] and use a measure of dispersion such as the variance to quantify uncertainty. Essentially, this means we rely on model disagreement as a metric for quantifying uncertainty. A high variance signifies numerous parameter configurations that yield equally low training loss ℓ, yet significantly differ in predicting outcomes for an unseen test point x.

As a consequence of this view, there is no reason to believe that GVI based $100(1-\alpha)\%$ prediction intervals for an unseen data point will have the pre-specified coverage. However, this shortcoming is not unique to GVI and is usually encountered in Bayesian models as well, owing to the difficulty in appropriately choosing the prior distribution and likelihood function [Fong and Holmes, 2021]. When it is important to generate accurate prediction intervals rather than merely confidence scores, it's feasible to convert raw uncertainties into well-calibrated prediction intervals using conformalisation [Vovk et al., 2005]. For conformal prediction, what matters primarily is the correct ranking of predictive outputs, where a larger confidence score corresponds to increased uncertainty. In my view, achieving such a ranking is typically the best outcome one can hope for and something that can be achieved successfully with GVI even in the absence of a Bayesian interpretation (cf. Chapter 4 & 5)

The GVI-problem is therefore of interest for researchers solely interested in minimising ℓ as well as for researchers who want to equip their algorithms with a measure of uncertainty. However, the true difficulty lies in finding the global minimiser Q^*. In the case where $D = \mathrm{KL}$ and P has Lebesgue-density p, we know that the global minimiser Q^* has a Lebesgue-density q^* [Bissiri et al., 2016] given as

$$q^*(\theta) = \frac{1}{Z} \exp\big(-\ell(\theta)\big) p(\theta)^\lambda \tag{2.4}$$

[1] Technically speaking we use the push-forward measure $f_{\#Q^*}(x)$ to induce probability distribution over the output space.

where $Z := \int \exp\big(-\ell(\theta)\big)p(\theta)^\lambda \, d\theta$. We can therefore calculate Z analytically or—if this is not feasible which is usually the case—use MCMC-based procedures [Andrieu et al., 2003] to obtain samples from the global minimiser Q^*.

In the general case, where $D \neq \mathrm{KL}$, finding Q^* is more challenging, since we do not have access to a formula for q^* in the spirit of (2.4). The only feasible approach prior to our work in Chapter 5 was to parameterise a set of probability distribution $\mathcal{Q} := \{Q_\nu \,:\, \nu \in \Gamma\} \subset \mathcal{P}(\Theta)$, called *variational family*, and solve a finite dimensional optimisation problem over the Euclidean parameter space Γ, i.e. on searches for

$$\nu^* \in \arg\min \big\{ L(Q_\nu) \,:\, \nu \in \Gamma \big\} \tag{2.5}$$

and uses Q_{ν^*} as approximation for Q^*. An approach that we refer to as *parameterised*-GVI and which has some obvious shortcomings (cf. Chapter 5).

2.2 Parameterised Generalised Variational Inference in Infinite Dimensions

At first glance, the GVI-problem in (2.1) extends naturally to the case where Θ is infinite. In fact, the conceptual arguments in favor of GVI outlined in the previous section carry over mutatis mutandis. However, there are several non-trivial questions for infinite-dimensional Θ: What space Θ should we choose? What is the relationship between the Bayesian posterior and the global minimiser Q^*? How can we chose a variational family $\mathcal{Q}$ leading to tractable GVI objective?

The first method deploying variational inference for an infinite-dimensional functional parameter space Θ is described in the seminal paper by Titsias [2009a] in the context of Gaussian process regression. However, even though Titsias [2009a] model is developed for the purpose of approximating the Gaussian process posterior distribution, there is no explicit mentioning of the underlying infinite-dimensional parameter space. In fact, in his model variational inference is only performed to approximate the Gaussian process posterior at a finite set of points, the data points and inducing points, effectively rendering the model a finite-dimensional one. This method is famously known as sparse variational Gaussian process (SVGP).

Matthews et al. [2016] later discovered that identical equations can be derived from a model that is truly formulated in the function space. In order to distinguish the situation notationally, we will from now on write f for the functional parameter instead of θ. Similarly, we denote the parameter space with E instead of Θ from now on. We can express the insights from Matthews et al. [2016] within the GVI-framework

laid out in section 2.1 as follows: SVGP performs parameterised GVI with loss L given as

$$L(Q) = \int \ell(f)\, dQ(f) + \mathrm{KL}(Q, P), \tag{2.6}$$

where $\ell(f) := \log p(y|f) := \sum_{n=1}^{n} \log \mathcal{N}\big(y_n \,|\, f(x_n), \sigma^2\big)$ with observation noise $\sigma^2 > 0$. The variational family is given as $\mathcal{Q} := \big\{GP(m_\nu, k_\nu) \,:\, \nu \in \Gamma\big\}$ where

$$\Gamma = \big\{\nu = (\mu, \Sigma, Z) \,:\, \mu \in \mathbb{R}^M,\, \Sigma \in \mathbb{R}^{M\times M} \text{ sym. and pos. definite},\, Z \in \mathcal{X}^M\big\} \tag{2.7}$$

with

$$m_\nu(x) := k_Z(x)^\top k_{ZZ}^{-1} \mu, \tag{2.8}$$

$$k_\nu(x, x') := k(x, x') - k_Z(x)^\top k_{ZZ}^{-1} k_Z(x') + k_Z(x)^\top k_{ZZ}^{-1} \Sigma k_{ZZ}^{-1} k_Z(x'). \tag{2.9}$$

for $x, x' \in \mathcal{X}$ and $Z \in \mathcal{X}^M$ the so called *inducing points*. The prior measure P is the Gaussian measure canonically associated with the Gaussian process $F \sim GP(0, k)$ with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which means that P is characterised by satisfying $\pi_D \# P = \mathcal{N}\big(0, k(D, D)\big)$ for any $D \in \mathcal{X}^J$ and any $J \in \mathbb{N}$ where $\#$ denotes the push-forward.

At first glance, this choice of variational family appears quite arbitrary but it is of paramount importance since it ensures that the KL-divergence between the two Gaussian measures $Q \in \mathcal{Q}$ and P on the function-space E reduces to a finite-dimensional KL-divergence between Gaussian measures on $\mathbb{R}^M$, i.e.

$$\mathrm{KL}(Q, P) = \mathrm{KL}\left(\pi_Z \# Q, \pi_Z \# P\right) = \mathrm{KL}\left(\mathcal{N}(\mu, \Sigma),\, \mathcal{N}\big(0, k_{ZZ}\big)\right) \tag{2.10}$$

for all $Q \in \mathcal{Q}$. This reduction to finite dimension is what makes variational inference in the function space tractable.

In Matthews et al. [2016], the parameter space E is chosen to be the space of all functions from $\mathcal{X}$ to $\mathbb{R}$, denoted as $\mathcal{F}(\mathcal{X}, \mathbb{R})$. However, Matthews et al. [2016] explicitly assumes, for the sake of applying Bayes' theorem, that the functional parameter space is a Polish space, which $\mathcal{F}(\mathcal{X}, \mathbb{R})$ only satisfies in the trivial case where $\mathcal{X}$ is countable [Munkres, 2019, Chapter 21, Example 2].

This (minor) inconsistency in the otherwise excellent paper by Matthews et al. [2016], led me down a quest for the *right* functional parameter space E that is consistent with Matthews' formalism and satisfies the requirement of being a Polish space.

The idea which is at the heart of my dissertation and foundational for the rest of this work is to replace

Gaussian processes with Gaussian random elements that take values in some abstract separable Banach space E[2]. The variational framework can then be rigorously reformulated for Gaussian random elements (cf. Chapter 3). In the special case where E is the space of continuous functions $C(\mathcal{X}, \mathbb{R})$, one retrieves Matthews et al. [2016] and therefore ultimately also the classical variational GP framework of Titsias [2009a].

The Gaussian random element approach was furthermore perfectly tailored to develop the theory for generalised variational inference in infinite dimensions (cf. Chapter 4). This extension of GVI led to some very practical benefits: The derivation of Gaussian Wasserstein inference, a method for function space inference with access to a tractable GVI loss for arbitrary variational mean and covariance function. This should be contrasted with SVGP where the requirements for a tractable KL-divergence between stochastic processes in (2.10) places strong constraints on the form of the variational mean and variational kernel (cf. (2.8) and (2.9)). The additional flexibility allows us to leverage deep learning architectures and surpassed SVGP on various benchmark data sets (cf. Table 4.1) showcasing the benefits of the Gaussian random element view and GVI for function space inference.

2.3 Generalised Variational Inference via the Wasserstein Gradient Flow

Although parameterised GVI methods have demonstrated remarkable success in providing approximate solutions to the GVI-problem [Blei et al., 2017], there is a rather obvious shortcoming inherent to all parameterised approaches. The GVI loss L in (2.1) is convex, as long as the regulariser $D(\cdot, P)$ is convex for the fixed reference measure $P \in \mathcal{P}(\Theta)$. However, the resulting optimisation problem for the variational parameters is non-convex in all but trivial cases, making it impossible to obtain any theoretical guarantees for GVI-procedures. From a mathematical point of view, it is therefore tempting to sidestep parametrisation altogether, and operate directly on the space of probability measures. The idea of performing gradient descent directly in $\mathcal{P}(\Omega)$ may seem impossible at first but there is in fact a rich mathematical literature on how notions of gradient flows can be generalised to arbitrary metric spaces [Ambrosio et al., 2005]. When the underlying metric space is given as the 2-Wasserstein space, it is common to refer to the associated gradient flow as *Wasserstein gradient flow*.

In Chapter 5 we explore how the Wasserstein gradient flow can be leveraged to design algorithms that perform gradient descent directly in the space of probability measures with finite second moment, denoted $\mathcal{P}_2(\Theta)$, where Θ is finite-dimensional. This allows us to contrast the algorithmic effects of different regularisers for GVI and serves as unifying principle for sampling approaches as diverse as deep ensembles,

[2]Every separable Banach space is a Polish space [Kechris, 2012, Chapter 3, Example 4]

the unadjasted Langevin algorithm and a system of repulsive interacting particles. We furthermore were able to provide the first asymptotic convergence guarantees for GVI with maximum mean discrepancy based regularisation. The Wasserstein gradient flow has previously been employed in machine learning for the purpose of solving optimisation problems on the space of probability measures [Arbel et al., 2019, Salim et al., 2020, Korba et al., 2021]. However, to the best of our knowledge we were the first to use it in the context of *true* GVI, i.e. when $D \neq \mathrm{KL}$.

The last chapter is the synthesis of the two big ideas in this thesis: we perform gradient descent for L in (2.1) (this time $D = \mathrm{KL}$) in the infinite-dimensional space of probability measures $\mathcal{P}_2(E)$ where E itself is an infinite-dimensional Hilbert space. The WGF leads us to the infinite-dimensional Langevin equation [Hairer et al., 2007a] and allows us to develop a novel method for approximately sampling Bayes posteriors that naturally carries over for non-Gaussian likelihoods and even multimodal functional posteriors. In particular, the derivation of the functional Wasserstein gradient flow equations rely crucially on the variational inference framework of Chapter 3 and therefore provide further justification for the rigorous treatment of the functional parameter space E. The infinite-dimensional Langevin equation has been studied previously for the purpose of sampling in function space [Dashti and Stuart, 2013, Ottobre et al., 2016, Lim et al., 2023]. However, to the best of our knowledge, nobody has derived *projected Langevin sampling* nor were the previous works motivated by a desire to implement the WGF for the GVI-problem with a functional parameter space.

3 | Variational Gaussian Processes: A Functional Analysis View

This chapter is based on the following publication:

> **Veit D. Wild*** and George Wynne*. "Variational Gaussian Processes: A Functional Analysis View." *Artificial Intelligence and Statistics (AISTATS), 2022*

Abstract

Variational Gaussian process (GP) approximations have become a standard tool in fast GP inference. This technique requires a user to select variational features to increase efficiency. So far the common choices in the literature are disparate and lacking generality. We propose to view the GP as lying in a Banach space which then facilitates a unified perspective. This is used to understand the relationship between existing features and to draw a connection between kernel ridge regression and variational GP approximations.

3.1 Introduction

Gaussian processes (GPs) are a ubiquitous modelling paradigm within machine learning [Rasmussen and Williams, 2005]. They are random functions with the useful property that their pointwise evaluations form a multivariate Gaussian random vector. Within the Bayesian framework one may use a Gaussian process as a prior for an unknown function, condition on observed information and then a posterior over the unknown function is obtained. Examples of applications include regression, classification, reinforcement learning and optimisation. See Rasmussen and Williams [2005] for an introduction. The GP framework enjoys such popularity since it is flexible, interpretable, has closed form expressions in some scenarios and offers a degree of uncertainty quantification.

An issue of the GP framework is the naive computational cost $O(N^3)$ to perform predictions, where N is the number of observed data points. This is due to a matrix inversion term. Many methods have been proposed to reduce this computational cost to something more palatable, both through theoretical [Csató and Opper, 2002, Seeger et al., 2003, Snelson and Ghahramani, 2006,0, Titsias, 2009a] and computational innovations [Gardner et al., 2018, Wang et al., 2019].

The focus of this paper is the variational inference paradigm where the posterior GP is approximated by an element of a candidate family through an optimisation routine where distance from the posterior GP is measured with the Kullback-Leibler (KL) divergence. The common candidate family is a family of GPs formed by conditioning the prior on M surrogate *features*, not necessarily equal to the observed information, resulting in $O(NM^2)$ complexity rather than the aforementioned $O(N^3)$. For example, features could be point evaluations or values of inner products against some user chosen set of functions. Foundational papers regarding variational inference for GPs include Titsias [2009a], Matthews et al. [2016] and for a survey consult Leibfried et al. [2020].

Choosing features to condition on is often conducted with the aim of closed form, or at least easy to compute, expressions. Therefore the features chosen can be very dependent on the given GP of interest through the covariance kernel, mean function or space the GP takes values in. This has led to a heuristic and somewhat ad-hoc approach in the literature to deriving features. Indeed, there is little in the way of a unified view of the choice of features and how different choices relate to each other.

Contributions: We present a unified perspective of existing features used in variational Gaussian process approximations by embracing the fact that GPs can be viewed as Gaussian random elements in Banach spaces. This perspective reveals generalisations of, and equivalences between, commonly used variational features. In particular we generalise the derivation for the popular variational Fourier features [Hensman

et al., 2018]. This injects rigour and clarity into the features used. Finally, a connection to kernel ridge regression is made which clarifies the role that the variational features play in the posterior approximation.

Existing work: The commonly employed framework of variational GP approximation was derived by Titsias [2009a]. Two other types of features we focus on are inter-domain [Lázaro-Gredilla and Figueiras-Vidal, 2009] and Fourier [Hensman et al., 2018]. A formalism of the use of Kullback-Leibler divergence over infinite dimensions, a key part of the variational GP methodology, was clarified by Matthews et al. [2016]. The Fourier features have recently been combined with spherical harmonics [Dutordoir et al., 2020] and rough path theory [Lemercier et al., 2021]. An appeal to Gaussian measures is made in Cheng and Boots [2016,0], Salimbeni et al. [2018]. The Gaussian process is associated with a Gaussian measure in an appropriate RKHS . This *dual* formulation is then used to derive more efficient variational GP approximations. However, although Gaussian measures are referenced in their work this avenue is not consistently pursued which separates their line of work from ours.

3.2 Gaussian Processes and Gaussian Random Elements

This section introduces notation and the necessary mathematical objects that will be relevant throughout the paper. We furthermore highlight problems with the Gaussian process formalism when it comes to variational inference.

3.2.1 Preliminaries and Notation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a the underlying probability space on which all random quantities are defined. Let $\mathcal{X}$ be a set. A family of random variables $G\colon \mathcal{X} \times \Omega \to \mathbb{R}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ is called a random process. Let $G(x)$ denote the random variable $G(x, \cdot)$. A random process is called a *Gaussian process* (GP) if for every $N \in \mathbb{N}$ and $\{x_n\}_{n=1}^N \subset \mathcal{X}$ the random vector $(G(x_1), \dots, G(x_N))$ is Gaussian. A Gaussian process is entirely determined by its mean function $m\colon \mathcal{X} \to \mathbb{R}$ defined $m(x) \coloneqq \mathbb{E}[G(x)]$ and covariance function, also know as covariance kernel, $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined $k(x, x') \coloneqq \mathbb{E}[(G(x) - m(x))(G(x') - m(x'))]$. We denote the GP with mean m and covariance kernel k as $G \sim GP(m, k)$. For background on Gaussian processes consult the works by Rasmussen and Williams [2005], Lifshits [2012], Adler [1990].

Let E be a separable Banach space, $\mathcal{B}(E)$ the Borel σ-algebra and $\mathcal{P}(E)$ the set of Borel probability measures on E. We will always assume that E is separable without explicitly stating it every time. The dual of E is defined as $E^* \coloneqq \{x^* : E \to \mathbb{R} \,|\, x^* \text{ is linear and continuous}\}$ and for $x^* \in E^*, y \in E$ we write $(y, x^*)_E \coloneqq x^*(y)$ for the so called *dual pairing*. A mapping $F\colon \Omega \to E$ is called a *Gaussian random element* (GRE) if for every $x^* \in E^*$ the real valued random variable $x^*(F) = (F, x^*)_E$ is

Gaussian. Each GRE has an associated mean $m \in E$ which is uniquely characterised by satisfying $(m, x^*)_E = \mathbb{E}[(F, x^*)_E]$ for all $x^* \in E^*$ and covariance operator $C\colon E^* \to E$ uniquely characterised by satisfying $(Cx^*, y^*)_E = \mathrm{Cov}[(F, x^*)_E, (F, y^*)_E]$ for all $x^*, y^* \in E^*$. We denote $F \sim \mathcal{N}(m, C)$ for a GRE with mean m and covariance operator C. Note that for $E = \mathbb{R}^N$ this coincides with the standard definition for the normal distribution in $\mathbb{R}^N$. An excellent introduction to GREs can be found in Chapter 4 of Van Neerven [2008].

A measure $P \in \mathcal{P}(E)$ is called a *Gaussian measure* (GM) if for every $x^* \in E^*$ the pushforward measure $P^{x^*} \in \mathcal{P}(\mathbb{R})$ defined by $P^{x^*}(\cdot) := (P \circ (x^*)^{-1})(\cdot)$ is a Gaussian measure on $\mathcal{B}(\mathbb{R})$. As with random variables in $\mathbb{R}$ and probability measures on $\mathcal{B}(\mathbb{R})$ there is a one-to-one correspondence between GREs and GMs on E. GREs and GMs can be studied in far more generality, or indeed with more specificity, than E being a Banach space [Bogachev, 1998, Da Prato, 2006]. For a gentle introduction into Gaussian measures on Banach spaces see chapter 3 in Hairer [2009].

When a GP G satisfies $\mathbb{P}(\{\omega\colon G(\cdot, \omega) \in E\}) = 1$ we say that its sample paths lie almost surely in E. This has been studied for numerous common choices of E [Rajput and Cambanis, 1972, Rajput, 1972, Lukić and Beder, 2001] and one can then identify the GP with a GRE, or equivalently with a GM, over E. Throughout the rest of this paper we shall be dealing purely in terms of GREs and later specific examples equating these to GPs shall be given. The use of GREs facilitates a general view of variational inference and will be the vehicle of our results.

3.2.2 In Search of the Right Formalism

In Matthews et al. [2016], the stochastic process perspective is employed to demonstrate that the ELBO presented in Titsias [2009a] can be interpreted as minimizing the KL divergence between the true posterior process and a suitably defined variational family. In this section, we aim to carefully unpack the mathematical details, which will lead us to the conclusion that the appropriate mathematical framework for these arguments involves Gaussian random elements in Polish spaces.

In Section 3.3 of Matthews et al. [2016], Matthews notes that he is interested in a 'probability measure on sets of functions $f : \mathcal{X} \to \mathbb{R}$.' Furthermore, he states that 'the prior measure P [...] is assumed to be a Gaussian process.' Mathematically, the probability measure P defined through a Gaussian process is a mapping $P : (\mathbb{R}^{\mathcal{X}}, \mathcal{S}) \to [0, 1]$, where $\mathbb{R}^{\mathcal{X}} := \{f : \mathcal{X} \to \mathbb{R}\}$ is the vector space of all functions from $\mathcal{X}$ to $\mathbb{R}$, and $\mathcal{S}$ is the smallest σ-algebra that makes all pointwise evaluation functionals $\pi_x : \mathbb{R}^{\mathcal{X}} \to \mathbb{R},\ f \mapsto f(x)$ measurable.

The issue with this choice of function space is that the space $\mathbb{R}^{\mathcal{X}}$ is, in general, too large. Specifically, $\mathbb{R}^{\mathcal{X}}$

is not typically a Polish space, which is often regarded as a minimum requirement for probability theory in infinite dimensions. For instance, when applying Bayes' theorem in infinite-dimensional settings, it is desirable that the specification of the prior measure, in combination with the Markov kernel describing the likelihood, ensures the existence of a Markov kernel for the posterior measure. In fact, Ghosal and van der Vaart [2017, Section 1.3] consider the existence of such a Markov kernel for the posterior measure as a necessary condition to speak of a 'true posterior measure.' They also note that a sufficient condition for this is that the prior measure is defined on the Borel σ-algebra of a Polish space.

In Section 4.1, Matthews et al. [2016] implicitly acknowledges the necessity of this assumption by repeatedly assuming that the spaces in which he operates are Polish. As argued above, we believe this assumption should have been introduced earlier, in Section 3.3, when Bayes' Theorem is first presented, in order to avoid potential pathologies arising from posterior measures that are not Markov kernels.

Bringing these points together, Matthews et al. [2016] requires a prior measure on a Polish function space that is 'Gaussian'. A broad class of Polish function spaces are Banach spaces. As discussed in Section 3.2.1, Gaussianity is well-understood within Banach spaces. Gaussian random elements in Banach spaces, along with their corresponding Gaussian measures, therefore provide a foundation for constructing a rigorous theory of variational inference in infinite dimensions. This assumption is general enough to cover most practical cases while being concrete enough to enable tractable computations.

3.3 Gaussian Random Element Regression

In this section we outline Gaussian random element regression which is the random element view of standard Gaussian process regression. This view is commonly employed in areas such as Bayesian inverse problems [Stuart, 2010]. At first glance the framework may appear (indulgently) abstract. However, we believe this is the most natural framework to investigate variational GP approximation. There is an important example at the end of the section showing all that follows does in fact coincide with standard GP regression.

3.3.1 The Regression Model for GREs

Let F be a GRE in E with mean m and covariance operator C and denote by $P \in \mathcal{P}(\mathcal{X})$ its corresponding GM. Suppose we have observations $Y = \{Y_n\}_{n=1}^N$ which are the image of F under some $\{D_n\}_{n=1}^N \subset E^*$, corrupted by independent scalar Gaussian noise

$$Y_n = (F, D_n)_E + \epsilon_n,$$

where $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$ independently for $n = 1, \ldots, N$. This can be equivalently expressed in perhaps more familiar notation as the probability density function (pdf) of Y given $F = f$ is

$$p(y|F = f) := \mathcal{N}\big(y|(f, D)_E, \sigma^2 I_N\big),$$

for $y \in \mathbb{R}^N, f \in E$, $(f, D)_E := \big((f, D_n)_E\big)_{n=1,\ldots,N}$ and $\mathcal{N}\big(\,\cdot\,|\mu, \Sigma\big)$ denotes the pdf of a Gaussian distribution on $\mathbb{R}^N$ with mean vector $\mu \in \mathbb{R}^N$ and covariance matrix $\Sigma \in \mathbb{R}^{N\times N}$.

In the Bayesian paradigm, one updates their beliefs about F after observing $Y := (Y_1, \ldots, Y_N)$ by combining the prior $F \sim \mathcal{N}(0, C)$ with the likelihood $p(y|F = f)$ to form a posterior. This can be a delicate task since E could be infinite dimensional. However, since in our scenario E is a Banach space and the measures corresponding to $Y|F = f$ are all dominated by the Lebesgue measure on $\mathbb{R}^N$ with a jointly measurable map $(y, f) \in \mathbb{R}^N \times E \to p(y|f) \in \mathbb{R}$, an infinite dimensional version of Bayes theorem applies [Ghosal and van der Vaart, 2017, Chapter 1.3].

It states that a regular version [Klenke, 2013, chapter 8.3] of the posterior measure exists, denoted $P^{F|Y} : \mathbb{R}^N \times \mathcal{B}(E) \to [0, \infty)$, $(y, A) \mapsto P^{F|Y=y}(A)$ and the measure $P^{F|Y=y}$ on $\mathcal{B}(E)$, which is the posterior measure of F given $Y = y$, is dominated by the prior measure P for any $y \in \mathbb{R}^N$ with Radon-Nikodym density $\frac{p(y|f)}{p(y)}$.

What all these technicalities really mean is that for $A \in \mathcal{B}(E)$

$$P^{F|Y=y}(A) = \int_A \frac{p(y|f)}{p(y)} dP(f), \tag{3.1}$$

where

$$\begin{aligned} p(y) &= \int_E p(y|F = f) dP(f) \\ &= \mathcal{N}\big(y|(m, D)_E, C_{DD} + \sigma^2 I_N\big) \end{aligned}$$

with

$$\big((m, D)_E\big)_n := (m, D_n) \tag{3.2}$$

$$(C_{DD})_{n,n'} := (CD_n, D_{n'})_E, \tag{3.3}$$

for all $n, n' = 1, \ldots, N$.

This posterior measure $P^{F|Y=y}$ is a GM since it is formed from the Gaussian likelihood $p(y|F = f)$ and

Gaussian prior $F \sim \mathcal{N}(m, C)$ (details in supplementary material section A.1). Denote the mean and covariance operator of $P^{F|Y=y}$ by $\widetilde{m}, \widetilde{C}$ respectively. As is usually the case with Bayesian techniques, the user is often not interested in the posterior measure itself but its pushforward through some prediction operation.

We focus on the case of two linear maps $T, T' \in E^*$ since the general case of $S \in \mathbb{N}$ elements can be handled analogously. The posterior mean $\widetilde{m} \in E$ satisfies

$$\big(\widetilde{m}, T\big)_E = (m, T)_E + C_{TD}(C_{DD} + \sigma^2 I_N)^{-1} y, \tag{3.4}$$

for any $T \in E^*$ and the posterior covariance operator $\widetilde{C} : E^* \to E$ satisfies

$$(\widetilde{C}T, T')_E = (CT, T')_E - C_{TD}(C_{DD} + \sigma^2 I_N)^{-1} C_{DT'}, \tag{3.5}$$

for all $T, T' \in E^*$, where C_{DD} as in (3.3) and $(C_{TD})_{1,n} \coloneqq (CT, D_n)_E$ for $n = 1, \dots, N$ and $C_{DT'} = C_{T'D}^\top \in \mathbb{R}^{N \times 1}$. The proof for these statements is given in section A.1 of the supplementary materials.

In summary, given a prior GRE and some observed values Y via some maps $\{D_n\}_{n=1}^N \subset E^*$ we can get a Gaussian posterior measure $P^{F|Y=y}$ for any $y \in \mathbb{R}^N$ on E. Two, or equivalently finitely many, linear functionals $T, T' \in E^*$ of F under the posterior will follow a multivariate Gaussian distribution and one can use (3.4) and (3.5) to calculate the mean vector and the covariance matrix.

3.3.2 Example: GPs with Continuous Paths

Before we give the promised example that links GP regression to GRE regression, we need to introduce some key results from functional analysis.

Let $\mathcal{X} \subset \mathbb{R}^D$ be compact with Borel σ-algebra $\mathcal{B}(\mathcal{X})$ and $C(\mathcal{X}, \mathbb{R})$ the space of continuous functions from $\mathcal{X}$ to $\mathbb{R}$ equipped with the standard supremum norm. Note that $C(\mathcal{X}, \mathbb{R})$ is a Banach space as long as $\mathcal{X}$ is compact. Denote by $R(\mathcal{X})$ the space of finite regular signed measures over $\mathcal{X}$ equipped with total variation norm [Rao and Rao, 1983, Section 2.4]. The Riesz-Markov theorem [Rao and Rao, 1983, Royden and Fitzpatrick, 2010, Corollary 4.7.6] states that $C(\mathcal{X}, \mathbb{R})^* = R(\mathcal{X})$ in the sense that each $\mu \in R(\mathcal{X})$ gives an element of $C(\mathcal{X}, \mathbb{R})^*$ via $f \mapsto \int_{\mathcal{X}} f(x) d\mu(x)$ and for every $T \in C(\mathcal{X}, \mathbb{R})^*$ there exists a unique $\mu \in R(\mathcal{X})$ such that $Tf = \int_{\mathcal{X}} f(x) d\mu(x)$. For example the pointwise evaluation map $\pi_x(f) = f(x)$ corresponds to the Dirac measure δ_x based at x for which we write $\pi_x = \delta_x$. For a covariance operator $C \colon C(\mathcal{X}, \mathbb{R})^* \to C(\mathcal{X}, \mathbb{R})$ we set $C\mu$ to be CT where T is the unique element of $C(\mathcal{X}, \mathbb{R})^*$ such that $Tf = \int_{\mathcal{X}} f(x) d\mu(x)$.

We now establish the connection between GREs and GPs when $E = C(\mathcal{X}, \mathbb{R})$. Let $G \sim GP(0, k)$ be a Gaussian process over a compact subset $\mathcal{X} \subset \mathbb{R}^d$ with kernel k and zero mean. Zero mean is for simplicity, non-zero can be handled straightforwardly.

Assume the GP has paths in $C(\mathcal{X}, \mathbb{R})$ with probability one. A standard result which provide a sufficient condition is the Kolmogorov continuity theorem [Øksendal, 2003, Theorem 2.2.3], if k is translation invariant then a condition regarding the decay of k is provided by Adler and Taylor [2007, Corollary 1.5.5] and a condition regarding the spectral measure [Rasmussen and Williams, 2005, Chapter 4.2.1] of k by Adler and Taylor [2007, Page 22].

Following Lifshits [2012, Example 2.4], see also Rajput and Cambanis [1972], any GP with almost surely continuous paths can be identified with a GRE F taking values in $E = C(\mathcal{X}, \mathbb{R})$ [1] , denote the corresponding GM by P. The covariance operator C of F is given as

$$C\nu(\cdot) = \int_{\mathcal{X}} k(\cdot, x')d\nu(x') \tag{3.6}$$

$$(C\nu, \mu)_E = \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, x')d\nu(x')d\mu(x), \tag{3.7}$$

for $\mu, \nu \in R(\mathcal{X})$. Using the identification of pointwise evaluation and Dirac measures mentioned above

$$\begin{aligned}
\mathrm{Cov}_P[F(x), F(x')] &= \mathrm{Cov}[(F, \delta_x)_E, (F, \delta_{x'})_E] \\
&= (C\delta_x, \delta_{x'})_E \\
&= \int \int k(t, t')d\delta_x(t)d\delta_{x'}(t') \\
&= k(x, x'),
\end{aligned}$$

for any $x, x' \in \mathcal{X}$ as expected.

In standard GP regression one observes corrupted pointwise information about the unknown function at a collection of points $X = \{x_n\}_{n=1}^N \subset \mathcal{X}$. So in the notation of the previous subsection $D_n = \delta_{x_n}$ is the map through which our observations are viewed.

Suppose we want to make a prediction at two new points $x, x' \in \mathcal{X}$. This corresponds to the measures $T = \delta_x$ and $T' = \delta_{x'}$ and we know that $\big(F(x), F(x')\big)|Y = y$ is multivariate Gaussian. From (3.4) we

[1] The identification discussed in Example 2.4 in Lifshits [2012] is as follows: Let $G : \Omega \times \mathcal{X} \to \mathbb{R}$ be a Gaussian process with kernel k, which is regular enough such that $G(\omega, \cdot)$ is continuos for $\mathbb{P}$-almost every $\omega \in \Omega$. Then $F : \Omega \to C(\mathcal{X}, \mathbb{R})$ defined via $F(\omega) := G(\omega, \cdot)$ is a GRE in $C(\mathcal{X}, \mathbb{R})$.

calculate the mean as

$$\widetilde{m}(x) = (\widetilde{m}, \delta_x)_E = k_{xX}(k_{XX} + \sigma^2 I_N)^{-1} y,$$

and similarly for $m(x')$. Furthermore the covariance between $F(x)$ and $F(x')$ under the posterior is given by formula (3.5) as

$$(\widetilde{C}\delta_x, \delta_{x'})_E = k(x, x') - k_{xX}(k_{XX} + \sigma^2 I_N)^{-1} k_{Xx'},$$

where k_{XX} is the matrix with n, n'-th entry $k(x_n, x_{n'})$ and $k_{xX} = (k(x, x_1), \ldots, k(x, x_N)) = k_{Xx}^\top$. This is the standard formula for the posterior mean and covariance of a GP given noisy pointwise observations [Rasmussen and Williams, 2005].

In summary, GRE regression on $E = C(\mathcal{X}, \mathbb{R})$ with observation functionals $D_n = \delta_{x_n}$, $n = 1, \ldots, N$ recovers standard GPR.

3.4 Variational Inference for Gaussian Random Elements

The posterior expressions (3.4) and (3.5) can have high computational cost since the matrix inverse term has naive cost $O(N^3)$. To avoid this cost the variational approximation paradigm is often used where $P^{F|Y=y}$ is approximated by selecting a measure in a candidate family $\mathcal{Q} \subset \mathcal{P}(E)$ that is optimal according to some divergence. The earliest works on this method are Titsias [2009a], Hensman et al. [2013] and, as discussed in the introduction, this area has received a lot of attention and innovation in the GP community recently [Hensman et al., 2018, Dutordoir et al., 2020, Lemercier et al., 2021].

We will now describe variational inference for Gaussian random elements in an abstract Banach space E. Much is owed to the work of Matthews et al. [2016], which formulated the important equations in the context of Gaussian processes. The following presentation applies in generality of Banach spaces which is the most general derivation the authors are aware of.

The following are desired properties of the family $\mathcal{Q}$ of candidates to approximate the posterior

1. Predictions involving any $Q \in \mathcal{Q}$ must be computationally tractable and less expensive than the true posterior.
2. $\mathcal{Q}$ contains measures that give a good approximation for the true posterior $P^{F|Y=y}$.
3. A measure of *closeness* between each $Q \in \mathcal{Q}$ and $P^{F|Y=y}$ must be tractable and cheap to evaluate.

Variational family: The idea in the construction of $\mathcal{Q}$ is to parameterise certain features of the target

posterior with a multivariate Gaussian, the hope being that these features will represent the target posterior well even though there may be less features than the number of points observed.

First choose $M \in \mathbb{N}$ elements from the dual $\{L_m\}_{m=1}^M \subset E^*$, the *features*, and set $L = (L_1, \ldots, L_M)$, $L\colon E \to \mathbb{R}^M$. Denote $U_m := (F, L_m)_E$, $m = 1, \ldots, M$ and $U := (U_1, \ldots, U_M)$. Define $Q^L := \mathcal{N}(\mu, \Sigma) \in \mathcal{P}(\mathbb{R}^M)$ for some mean vector $\mu \in \mathbb{R}^M$ and covariance matrix $\Sigma \in \mathbb{R}^{M\times M}$. Starting with a Q^L of this form we obtain a member of the approximating family Q as

$$Q(A) = \int_A \left(\frac{dQ^L}{dP^L} \circ L\right)(f) dP(f)$$

where $A \in \mathcal{B}(E)$ and dQ^L/dP^L is the Radon-Nikodym derivative of Q^L with respect to P^L and P^L is the law of U under the prior equal to $\mathcal{N}((m, L)_E, C_{LL})$. The idea is that Q^L dictates the behaviour of Q on the features L.

While this formulation of a candidate $Q \in \mathcal{Q}$ may seem obtuse, in Theorem A.1.1 in the Supplement it is shown

$$Q(A) = \int_{\mathbb{R}^M} \mathbb{P}\big(F \in A \big| U = u\big) dQ^L(u),$$

which is used to deduce that each $Q \in \mathcal{Q}$ is a GM with mean m_Q

$$(m_Q, T)_E = (m, T)_E + C_{LL}^{-1}\big(\mu - (m, L)_E\big) C_{LT} \tag{3.8}$$

for all $T \in E^*$ and covariance operator C_Q

$$\big(C_Q T, T'\big)_E = (CT, T')_E + C_{TL} C_{LL}^{-1} (\Sigma - C_{LL}) C_{LL}^{-1} C_{LT'} \tag{3.9}$$

for all $T, T' \in E^*$.

The variational parameters of this family are μ, Σ and potentially parameters that appear in the specification of the inducing features L. For ease of notation denote all of these parameters by η and the Q corresponding to this choice by Q_η.

Measure of closeness: The Kullback-Leibler (KL) divergence is the measure of closeness employed. For $P, Q \in \mathcal{P}(E)$ with Q absolutely continuous with respect to P, denoted $Q \ll P$

$$KL(Q, P) = \int_E \log\left(\frac{dQ}{dP}(f)\right) dQ(f),$$

and $KL(Q, P)$ is infinite if Q is not absolutely continuous with respect to P. The variational parameter $\eta \in \Gamma$ is then selected by minimising the KL

$$\eta^* \in \arg\min_{\eta} KL\big(Q_\eta, P^{F|Y=y}\big).$$

After η^* has been determined, the posterior is approximated with Q_{η^*} which we denote by Q^* for ease of notation.

The choice of $\mathcal{Q}$ means one may rewrite the KL as

$$KL(Q, P^{F|Y=y}) = KL(Q, P) - \mathbb{E}_Q[\log p(y|F)] + \log p(y)$$

where $\mathbb{E}_Q[\log p(y|F)] := \int \log p(y|F = f) dQ(f)$, see Theorem A.2.2 in the Supplement.

Optimisation: Optimisation with respect to KL is performed by optimising the evidence based lower bound (ELBO) defined $\mathcal{L} := -KL(Q, P) + \mathbb{E}_Q[\log p(y|F)]$.

The user can now optimise the parameters η, which we recall are μ, Σ, analytically to obtain the optimal $Q^* \in \mathcal{Q}$ within the candidate family. The parameters μ, Σ have a closed form expression and cost $\mathcal{O}(NM^2 + M^3)$ [Titsias, 2009a]. These optimal choices for μ, Σ, given fixed L, are given in Theorem A.2.2 in the Supplement. The resulting optimal mean and covariance operators, denoted m_{Q^*} and C_{Q^*} satisfy

$$\begin{aligned} (m_{Q^*}, T)_E &= C_{TL}\big(\sigma^2 C_{LL} + C_{LD}C_{DL}\big)^{-1} C_{LD} y \\ (C_{Q^*}T, T')_E &= (CT, T')_E - C_{TL}C_{LL}^{-1}C_{LT'} + C_{TL}\big(C_{LL} + \frac{1}{\sigma^2}C_{LD}C_{DL}\big)^{-1}\big)C_{LT'}, \end{aligned} \tag{3.10}$$

for all $T, T' \in E^*$. See Theorem A.2.2 in the Supplement for a proof.

Alternatively, a user could numerically optimise η using a factorisation of $\mathcal{L}$ over N to make use of batch size optimisation [Hensman et al., 2013]. This leads to complexity $\mathcal{O}(N_B M^2 + M^3)$, where $N_B \in \mathbb{N}$ is the batch size.

The factorised ELBO is normally used for really large data sets and in this case the bottleneck is the inversion of C_{LL} which causes the $\mathcal{O}(M^3)$ complexity term. Therefore it is vital for practitioners to choose L which result in C_{LL} being easy to invert, for example making C_{LL} be diagonal.

The next section investigates common choices of L in the literature and derives a unifying perspective using GREs, crucial for understanding the different choices.

3.5 Functional Analysis View on Inducing Features

In this section several variational approaches are recovered within the GRE framework. The goal is obtaining a unified perspective and greater generality of the derivations.

As described in Section 3.3.2, the starting point for our analysis is a GRE $F \sim \mathcal{N}(m, C)$ in $E = C(\mathcal{X}, \mathbb{R})$ whose corresponding measure on E is denoted P. This GRE is then conditioned upon corrupted pointwise observations $\{x_n\}_{n=1}^N \subset \mathcal{X}$ such that $Y_n = (F, \delta_{x_n})_E + \epsilon_n$ with $\epsilon_1, \ldots, \epsilon_N \sim \mathcal{N}(0, \sigma^2)$.

Different choices of features $\{L_m\}_{m=1}^M \subset C(\mathcal{X}, \mathbb{R})^*$ shall lead to the original inducing point approach [Titsias, 2009a], inter-domain features [Lázaro-Gredilla and Figueiras-Vidal, 2009] and variational Fourier features [Hensman et al., 2018].

While the first two examples appear pedestrian the third crucially relies upon the GRE to reveal how Fourier features actually behave and under what conditions they are valid, greatly expanding their scope beyond the example in Hensman et al. [2018].

Only the covariances C_{LL}, C_{LT} of the features are derived since this is all that is needed to compute the variational mean m_Q and covariance operator C_Q of the approximating Q, see (3.8) and (3.9).

The prediction map T will always be a single point evaluation at an arbitrary point $x \in \mathcal{X}$. The case of other choices of T, in particular point evaluation at multiple locations, is straightforward. The term C_{LL} is the bottleneck term in the computation, as discussed in the previous section.

3.5.1 Background in Functional Analysis

In this section we introduce some basic terminology and results from functional analysis. The reader unfamiliar with these tools is referred to Chapter 13 of Royden and Fitzpatrick [2010] for additional information.

Let E be a Banach space and $\mathcal{X} \subset \mathbb{R}^D$ be compact[2]. Typical examples of Banach spaces E are the space of continuous functions $C(\mathcal{X}, \mathbb{R})$ endowed with the supremums norm and the space $L^2(\mathcal{X}, \mathbb{R})$ of equivalence classes of square integral functions, which is even a Hilbert space.

Linear operators A linear map $L : E \to W$ (typically called operator) between two Banach spaces $\big(E, ||\cdot||_E\big)$ and $\big(W, ||\cdot||_W\big)$ is called bounded iff

$$||L|| := \sup_{v \in E} \frac{||Lv||_W}{||v||_E}$$

[2]The compactness of $\mathcal{X}$ is required to make $C(\mathcal{X}, \mathbb{R})$ a Banach space, but is not necessary for the Hilbert space structure of $\mathcal{L}^2(\mathcal{X}, \mathbb{R})$.

is finite. For every operator $L : E \to W$ we define the adjoint operator $L^* : W^* \to E^*$ via $L^*(\phi)(v) := \phi(Lv)$ for all $\phi \in W^*$, $v \in E$. This can be equivalently expressed with the help of the dual pairing as $\big(v, L^*\phi\big)_E = \big(Lv, \phi\big)_W$. If E and W are Hilbert spaces we can interpret the dual pairing as the respective Hilbert space inner-products via the Riesz-representation theorem.

Kernel integral operator Let k be a (continuous) kernel on $\mathcal{X}$ and define $T_k : L^2(\mathcal{X}, \mathbb{R}) \to L^2(\mathcal{X}, \mathbb{R})$ as $T_k f(x) = \int_{\mathcal{X}} k(x, x')f(x')dx'$. The operator T_k is called kernel (integral) operator and is well-studied in machine learning [cp. Steinwart and Scovel, 2012]. It can be easily shown that T_k is bounded and self-adjoint, i.e. $T_k^* = T_k$.

Linear transformation of Gaussian Measures Let $L : E \to W$ be a bounded linear operator and $F \sim \mathcal{N}(m, C)$ be a GRE in E. Then $L(F)$ is a GRE in W with mean Lm and covariance operator LCL^* [Hairer, 2009, chapter 3.3].

3.5.2 Inducing Points

The inducing-points framework of Titsias [2009a] chooses the inducing features as pointwise evaluations $L_m = \pi_{z_m}$, which corresponds to the measure δ_{z_m} for some set of points $\{z_m\}_{m=1}^M \subset \mathcal{X}$.

Substituting this choice of L into (3.7)

$$\begin{aligned}
(C_{LL})_{mm'} &= \mathrm{Cov}_P(L_m F, L_{m'} F) \\
&= \mathrm{Cov}_P[(F, \delta_{z_m})_E, (F, \delta_{z_m})_E] \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, x') d\delta_{z_m}(x) \delta_{z_{m'}}(x') \\
&= k(z_m, z_{m'}),
\end{aligned}$$

and

$$\begin{aligned}
(C_{TL})_m &= \mathrm{Cov}_P(TF, L_m F) \\
&= \mathrm{Cov}_P[(F, \delta_x)_E, (F, \delta_{z_m})_E] \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} k(t, t') d\delta_x(t) \delta_{z_m}(t') = k(x, z_m).
\end{aligned}$$

Combining these calculations with (3.8), (3.9) we recover the same formula as seen in the original inducing point derivation [Titsias, 2009a].

3.5.3 Inter-domain Features

The natural space to realise inter-domain features [Lázaro-Gredilla and Figueiras-Vidal, 2009] is $L^2(\mathcal{X}, \mathbb{R})$ since they involve an inner product on $L^2(\mathcal{X}, \mathbb{R})$. To this end map the GRE F that takes values in $C(\mathcal{X}, \mathbb{R})$ into $L^2(\mathcal{X}, \mathbb{R})$ via the canonical embedding $\iota : C(\mathcal{X}, \mathbb{R}) \to L^2(\mathcal{X}, \mathbb{R})$, $f \mapsto f$. The adjoint operator of ι is given as $\iota^* : L^2(\mathcal{X}, \mathbb{R}) \to R(\mathcal{X})$, $f \mapsto \int_{(\cdot)} f(x)dx$, which can be easily verified. As explained in Section 3.3.2 the space $C(\mathcal{X}, \mathbb{R})^*$ can be identified with the space of signed Borel measures on $\mathcal{B}(\mathcal{X})$ which explains why $\iota^*(f)$ is an element of $R(\mathcal{X})$ and therefore accepts as input a set to integrate over.

It follows from $||f||_{L^2(\mathcal{X},\mathbb{R})} \leq \sqrt{\lambda(\mathcal{X})}||f||_\infty$, where $||\cdot||_\infty$ is the supremum norm and λ the Lebesgue measure, that ι is a bounded, linear operator from $C(\mathcal{X}, \mathbb{R})$ to $L^2(\mathcal{X}, \mathbb{R})$. We can use the result mentioned in Section 3.5.1 to conclude ιF is a GRE in $L^2(\mathcal{X}, \mathbb{R})$ with covariance operator $\iota C \iota^*$ given as

$$(\iota C \iota^*)(g) = C\big(\iota^*(g)\big) = \int_{\mathcal{X}} k(\cdot, x')g(x')dx', \tag{3.11}$$

where (3.11) is simply from the definition of the covariance operator of the GRE on $C(\mathcal{X}, \mathbb{R})$. By (3.11) $\iota C \iota^*$ coincides with the well-studied integral operator described in Section 3.5.1 and denoted as T_k.

Inter-domain features can be written in our notation as

$$L_m F = \langle \iota F, g_m \rangle_{L^2(\mathcal{X},\mathbb{R})} = \int_{\mathcal{X}} F(x) g_m(x) dx$$

for some collection $\{g_m\}_{m=1}^M \subset L^2(\mathcal{X}, \mathbb{R})$. Using the definition of an adjoint operator this is equal to $L_m F = (F, \iota^* g_m,)_E$ so $L_m = \iota^* g_m$.

Substituting into (3.7)

$$\begin{aligned}(C_{LL})_{mm'} &= \mathrm{Cov}_P(L_m F, L_{m'} F) \\ &= \mathrm{Cov}_P[(F, \iota^* g_m)_E, (F, \iota^* g_{m'})_E] \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, x') g_m(x) g_{m'}(x') dx dx', \end{aligned} \tag{3.12}$$

and

$$\begin{aligned}(C_{TL})_m &= \mathrm{Cov}_P(TF, L_m F) \\ &= \mathrm{Cov}_P[(F, \delta_x)_E, (F, \iota^* g_m)_E] \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} k(t, x') g_m(x') d\delta_x(t) dx' = T_k(x), \end{aligned} \tag{3.13}$$

which agrees with the original derivation of inter-domain features [Lázaro-Gredilla and Figueiras-Vidal, 2009].

3.5.4 Fourier Features

Fourier features are defined as a reproducing kernel Hilbert space (RKHS) inner product between F and trigonometric functions [Hensman et al., 2018].

An RKHS is a Hilbert space of functions that, as the name suggests, is associated with a kernel. Namely, given a kernel k the RKHS is the unique Hilbert space of functions mapping from $\mathcal{X}$ to $\mathbb{R}$, which we denote H_k, such that $k(\cdot, x) \in H_k$ for all $x \in \mathcal{X}$ and $\langle f, k(\cdot, x)\rangle_k = f(x)$ for all $f \in H_k$ and $x \in \mathcal{X}$, where $\langle \cdot, \cdot \rangle_k$ denotes the inner product on H_k. This latter property is called the reproducing property. For more on the theory of RKHS consult Berlinet and Thomas-Agnan [2004].

The idea of Fourier features is to observe an inner product not in $L^2(\mathcal{X}, \mathbb{R})$, as was done in inter-domain features, but instead to observe an inner product in H_k. Namely, Hensman et al. [2018] use features that would be written in our framework as $L_m F = \langle F, g_m \rangle_k$ where $\{g_m\}_{m=1}^M$ are the first M elements of the Fourier basis. However, it is well known that $F \notin H_k$ almost surely [Lukić and Beder, 2001] therefore the aforementioned choice of L *cannot* be used without extra justification.

Hensman et al. [2018] provided justification in the particular case when k is a Matérn kernel with certain parameters over $\mathcal{X} = \mathbb{R}$. An explicit form of the Matérn RKHS inner product is used and the core of the argument is the g_m are "very regular" in an appropriate sense to compensate for the way that almost surely F is not regular enough to be contained in H_k [Hensman et al., 2018, Section 3.3.1].

A rigorous justification is now given for these type of features. The derivation is more general than just the one-dimensional Matérn case and a condition on which functions g_m can be used instead of the Fourier basis is provided.

For the rest of this section we will assume that k is continuous. Recall the kernel integral operator T_k defined above, the square root $T_k^{1/2}$ is an isometric isomorphism between $L^2(\mathcal{X}, \mathbb{R})$ and H_k [Steinwart and Christmann, 2008, Theorem 4.51] meaning that $T_k^{1/2}(L^2(\mathcal{X}, \mathbb{R})) = H_k$ and

$$\langle T_k^{1/2} f, T_k^{1/2} g\rangle_k = \langle f, g\rangle_{L^2(\mathcal{X},\mathbb{R})}, \tag{3.14}$$

for all $f, g \in L^2(\mathcal{X}, \mathbb{R})$. The moral of this result is that $T_k^{1/2}$ bestows upon element of $L^2(\mathcal{X}, \mathbb{R})$ just enough regularity to be in H_k. This notion of adding regularity will tie into the notion of "very regular" employed by Hensman et al. [2018].

Define features $L_m = \iota^* f_m$ for some $\{f_m\}_{m=1}^M \subset L^2(\mathcal{X}, \mathbb{R})$ so

$$L_m F = (F, \iota^* f_m)_E = \langle \iota F, f_m \rangle_{L^2(\mathcal{X},\mathbb{R})}$$

where $\iota \colon C(\mathcal{X}, \mathbb{R}) \to L^2(\mathcal{X}, \mathbb{R})$ is the inclusion operator used in the previous subsection. Then

$$\begin{aligned} \text{“}\langle F, T_k f_m \rangle_k = \text{”} \langle T_k^{1/2} \iota F, T_k^{1/2} f_m \rangle_k & \qquad (3.15) \\ = \langle \iota F, f_m \rangle_{L^2(\mathcal{X},\mathbb{R})} & \qquad (3.16) \\ = (F, \iota^* f_m)_E = L_m F. & \qquad (3.17) \end{aligned}$$

The first expression is in quotes since $F \notin H_k$ almost surely [Lukić and Beder, 2001] so the expression has no real meaning. We include it though since if one were to suspend reality then the equality in quotes would be valid since $T_k^{1/2}$ is self-adjoint on H_k so it can be borrowed from $T_k f_m$. The second term in (3.15) is well defined since $T_k^{1/2}$ maps from $L^2(\mathcal{X}, \mathbb{R})$ to H_k. The move to (3.16) is facilitated by the isometry (3.14).

The main idea of what is happening in (3.15) is F is borrowing a $T_k^{1/2}$ from $T_k f_m$ to be able to live in the RKHS. This is happening explicitly in the calculations done by Hensman et al. [2018] in the Matérn case. Indeed, the way that $T_k f_m$ is the image of f_m under *two* applications of $T_k^{1/2}$, rather than just the standard one needed to be in the RKHS, is the mathematical explanation of the notion of “very regular” that was alluded to by Hensman et al. [2018] since it has had two portions of the regularity provided by $T_k^{1/2}$.

It is interesting to see that RKHS inner product feature (3.15) can be reduced to (3.17) which is simply using $T_k f_m$ as an inter-domain feature. So using the inter-domain formula for C_{LL} (3.12) gives

$$\begin{aligned} (C_{LL})_{mm'} &= \text{“}\mathrm{Cov}_P(\langle F, T_k f_m \rangle_k, \langle F, T_k f_{m'} \rangle_k)\text{”} \\ &= \mathrm{Cov}_P[(F, \iota^* f_m)_E, (F, \iota^* f_{m'})_E] \\ &= \langle T_k f_m, f_{m'} \rangle_{L^2(\mathcal{X},\mathbb{R})} \\ &= \langle T_k^{1/2} f_m, T_k^{1/2} f_{m'} \rangle_{L^2(\mathcal{X},\mathbb{R})} \qquad (3.18) \\ &= \langle T_k f_m, T_k f_{m'} \rangle_k, \qquad (3.19) \end{aligned}$$

where (3.18) from the fact that $T_k^{1/2}$ is self-adjoint considered as operator on $L^2(\mathcal{X}, \mathbb{R})$ and (3.19) is using the isometry between $L^2(\mathcal{X}, \mathbb{R})$ and H_k and. Similarly,

$$(C_{TL})_m = \text{“}\mathrm{Cov}_P(\langle F, k(x, \cdot) \rangle_k, \langle F, T_k f_m \rangle_k)\text{”}$$

$$= \mathrm{Cov}_P[(F, \delta_x)_E, (F, \iota^* f_m)_E]$$
$$= T_k f_m(x) \tag{3.20}$$

where (3.20) is using (3.13). The connection to Hensman et al. [2018] is revealed once one sets $g_m = T_k f_m$ where g_m are the features employed by Hensman et al. [2018]. In particular (3.19) and (3.20) are equal to Equation 61 and Equation 60, respectively, in Hensman et al. [2018].

As was done by Hensman et al. [2018] it makes sense from a practical point of view to define $T_k f_m$ without explicitly choosing f_m. Our derivation shows this may be performed without dependence on kernel parameters as long as g_m is in the range of T_k for all the kernel parameters that could be considered.

Our derivation is general enough to justify the use of Fourier features in Dutordoir et al. [2020], where zonal kernels in more than one input dimension are used. Their choice of g_m corresponds to eigenfunctions in the Mercer expansion of the kernel which is always in the image of T_k.

3.6 Gaussian Random Elements and the Nyström Method

In this section we demonstrate how the GRE framework reveals further connections between Gaussian process regression and the Nyström approximation for kernel Ridge regression (KRR). These connections have been known for a while and received some attention recently [Parzen, 1961, Wahba, 1990, Kanagawa et al., 2018, Wild et al., 2021]

Let k be a kernel, H_k the corresponding RKHS and $\{x_n, y_n\}_{n=1}^N \subset \mathcal{X} \times \mathbb{R}$ be paired observations. In Nyström KRR [Williams and Seeger, 2001] we seek to minimise the empirical risk over a finite dimensional subspace $\mathcal{M} \subset H_k$

$$\widehat{f} \coloneqq \operatorname*{argmin}_{f \in \mathcal{M}} \frac{1}{N} \sum_{n=1}^{N} \big(f(x_n) - y_n\big)^2 + \lambda ||f||_k^2, \tag{3.21}$$

where $\lambda > 0$ is a regularisation parameter and $\widehat{f}$ is called the Nyström approximation over $\mathcal{M}$.

The subspace $\mathcal{M}$ that is typically selected is $\mathrm{span}\big(\{k(\cdot, z_m)\}_{m=1}^M\big)$, where $\{z_m\}_{m=1}^M \subset \mathcal{X}$ are user chosen points, referred to as landmark points. Most research has focused on different sampling approaches for the landmark points to guarantee high quality approximations [Rudi et al., 2015, Musco and Musco, 2016, Li et al., 2016].

The GRE perspective facilitates a generalised view in which it becomes clear how the choice of features L in variational GRE regression, see Section 3.4, corresponds to the choice of subspace $\mathcal{M}$ in Nyström

KRR.

Theorem 3.6.1. *Let $F \sim \mathcal{N}(0, C)$ be a GRE in $E = C(\mathcal{X}, \mathbb{R})$ with covariance operator C as defined in (3.6) and assumed pointwise noisy data is observed as described in Section 3.3.2. For $\{\mu_m\}_{m=1}^M \subset R(\mathcal{X})$ let $L_m = \mu_m$ be the features used in the variational approximation. Set $\mathcal{M} = \{C\mu_m\}_{m=1}^M$ where $C\mu_m = \int k(\cdot, x')d\mu_m(x')$ as the approximating family in the Nyström approximation. Then for $\sigma^2 = N\lambda$ the Nyström KRR estimator $\widehat{f}$ in Section 3.6 is equal to the mean m_{Q^*}, given by (3.10), of the optimal Q^* from the variational family $\mathcal{Q}$.*

We give the proof and an explicit form of $\widehat{f}$ and m_{Q^*} in Section A.3 of the Supplement.

The GRE view reveals that the connecting element between the Nyström approximations and variational GPs is via the signed measures $\{\mu_m\}_{m=1}^M$. Knowledge of the signed measures either from inspecting the elements of the set $\mathcal{M}$ in the Nyström approximation or the projections L_m in variational GPs will allow to translate results from one field to another. This connection was previously only known for the special case of inducing inducing points $\mu_m := \delta_{z_m}$ outlined in Wild et al. [2021]. This opens the door to applying the theory of Nyström KRR error bounds to variational GPs to gain a better understanding of the latter's approximation properties. Vice versa, recent advances in variational GPR approaches, for example, variational Fourier features, could be leveraged in the context of KRR Nyström, as they simply correspond to a particular choice of $\mathcal{M}$.

3.7 Conclusion

We have outlined the GRE framework as a technical tool to provide a unified and generalising perspective to variational GP approximation. Along the way we have seen how existing choices of features, previously thought distinct, are in fact highly related and how they can be derived in wider settings than those currently employed. Finally, we related the posterior mean of variational GP approximations with a Nyström KRR approximation which offers a new lens to view the impact of different feature choices in variational GP approximations.

Statement of Authorship for joint/multi-authored papers for PGR thesis

Title of Paper	Variational Gaussian Processes: A Functional Analysis View
Publication Status	Published
Publication Details	Veit D. Wild* and George Wynne*. "Variational Gaussian Processes: A Functional Analysis View." Artificial Intelligence and Statistics (AISTATS), 2022.

Student Confirmation

Student Name:	Veit David Wild		
Contribution to the Paper	• Idea of Variational inference in function space with Gaussian random elements instead of GPs. • Derivation of the measure theoretic framework for KL posterior and ELBO. • Derivation of Theorem 1. • Writing a large part of the paper		
Signature	V.Wild	Date	28.01.2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Dino Sejdinovic		
Supervisor comments			
Signature	Dino Sejdinovic	Date	30 January 2024

This completed form should be included in the thesis, at the end of the relevant chapter.

4 | Generalised Variational Inference in Function Spaces: Gaussian Measures Meet Bayesian Deep Learning

This chapter is based on the following publication:

> **Veit D. Wild***, Robert Hu* and Dino Sejdinovic. "Generalized Variational Inference in Function Spaces: Gaussian Measures meet Bayesian Deep Learning." *Advances in Neural Information Processing Systems (NeurIPS), 2022*

Abstract

We develop a framework for generalised variational inference in infinite-dimensional function spaces and use it to construct a method termed Gaussian Wasserstein inference (GWI). GWI leverages the Wasserstein distance between Gaussian measures on the Hilbert space of square-integrable functions in order to determine a variational posterior using a tractable optimisation criterion. It avoids pathologies arising in standard variational function space inference. An exciting application of GWI is the ability to use deep neural networks in the variational parametrisation of GWI, combining their superior predictive performance with the principled uncertainty quantification analogous to that of Gaussian processes. The proposed method obtains state-of-the-art performance on several benchmark datasets.

4.1 Introduction

In the past decade, considerable effort has been invested in developing Bayesian deep learning approaches [Welling and Teh, 2011, Chen et al., 2014, Blundell et al., 2015, Gal and Ghahramani, 2016, Kendall and Gal, 2017, Ritter et al., 2018, Khan et al., 2018, Maddox et al., 2019]. There are at least two key advantages to Bayesian models. Firstly, Bayesian model averaging is known to improve predictive performance [Komaki, 1996] even in misspecified situations [Fushiki, 2005, Ramamoorthi et al., 2015]. The empirical success of methods such as deep ensembles [Lakshminarayanan et al., 2017] may be interpreted as compelling evidence for this claim [Wilson and Izmailov, 2020]. Secondly, Bayesian models provide the user with a predictive distribution for an unseen data point. This can be naturally leveraged to quantify posterior uncertainty.

Even though impressive progress has been made, there are problems that remain unresolved. The prior distribution for the unknown function is typically induced by a prior distribution over deep neural network weights (and biases). It is hard to interpret the inductive bias in a function space that is induced by such priors for weights and unclear how one might incorporate prior knowledge about the unknown function. Additionally, the resulting inference problem is extremely high-dimensional and requires approximation techniques that are either computationally expensive [Neal, 2012] or so crude that the approximate posterior may suffer from pathological behavior [Foong et al., 2020]. The difficulties of performing Bayesian inference for weights have led to the emergence of methods that approach the problem directly in function space [Ma et al., 2019, Sun et al., 2018, Rudner et al., 2020, Ma and Hernández-Lobato, 2021].

The theory of constructing prior distributions in function spaces is well developed and the most famous class of prior distributions are *Gaussian processes*. They have been commonly used for decades in the machine learning community to elicit interpretable functional priors and are known to have well-calibrated predictive uncertainties [Rasmussen and Williams, 2006].

In a separate thread of research, a new powerful inference framework called *Generalised Variational Inference* (GVI) has been recently developed [Knoblauch et al., 2019]. The authors argue that standard assumptions of Bayesian inference such as well-specified priors, well-specified likelihoods and infinite computing power are often violated in practice. They therefore propose a generalised view on Bayesian inference that takes these points into consideration. We extend the work of Knoblauch et al. [2019] to situations where no probability density functions for the prior exist and are thus able to use generalised variational inference in infinite-dimensional function spaces directly. We then specify both the prior and variational measures as Gaussian measures and measure their dissimilarity using the Wasserstein distance. This results in the method which we call *Gaussian Wasserstein Inference in Function Spaces* (GWI-FS).

An exciting application of our method is the ability to equip deep neural networks with uncertainty quantification using the framework analogous to that of Gaussian processes, resulting in a state-of-the-art method termed *GWI-net*. Our main contributions are:

- We create a general framework for inference in function space based on Gaussian measures on the space of square-integrable functions,
- We derive an objective function that can be expressed in terms of the *parameters of the Gaussian measures*,
- We derive a tractable approximation to our objective function that is valid for (almost) arbitrary kernels and mean functions,
- We demonstrate the utility of our method by obtaining state-of-the-art results on the UCI regression datasets and on Fashion MNIST and CIFAR 10[1].

4.2 Related Work

GWI-FS draws on the work developed in the Gaussian process literature, but can be used to equip traditional neural network architectures with uncertainty. We therefore give a brief overview of the relevant related methods in both the Bayesian neural network (BNNs) and Gaussian process community.

Bayesian neural networks Traditionally Bayesian neural networks have been assigned priors in weight space. The effects of various priors on inference and uncertainty quantification are still not well understood [Fortuin et al., 2021]. As the posterior (over weights) is intractable, sampling algorithms such as Hamiltonian Monte Carlo (HMC) were initially proposed Neal [2012]. Due to the unfavorable scaling properties of standard HMC which requires the full gradient, batch-size approximations of HMC evolved [Chen et al., 2014]. Another line of research exploits Langevin dynamics to generate posterior samples [Welling and Teh, 2011] in weight space.

Variational methods for BNNs in weight space In variational inference, the true posterior is approximated by a more tractable so-called *variational* distribution. The user specifies a class of approximate posterior measures and selects the best posterior approximation by maximising the so-called evidence lower bound (ELBO). The Bayes by Backprop [Blundell et al., 2015] method is one such variational mean-field approximation of the weight-space posterior. In variational dropout [Gal and Ghahramani, 2016], a specific approximation is chosen to reinterpret dropout [Srivastava et al., 2014] at test time as a variational procedure.

Variational methods for BNNs in function spaces Inference in weight space is challenging, as the

[1]Codebase: https://github.com/MrHuff/GWI

problem is typically high-dimensional and the posterior distribution over weights multi-modal. This led to a line of research in which inference algorithms are formulated in function spaces. Variational implicit processes [Ma et al., 2019] approximate the BNN posterior as a linear combination of draws from the prior. Functional-BNN [Sun et al., 2018] matches a BNN to a functional prior (for example a GP) and performs inference by optimising a functional Kullback-Leibler (KL) divergence exploiting score function estimators [Li and Turner, 2017, Shi et al., 2018]. Rudner et al. [2020] use a local approximation to the prior and variational posterior processes to obtain a tractable functional Kullback-Leibler divergence. Ma and Hernández-Lobato [2021] generalise the variational family in Ma et al. [2019] and obtain a more scalable procedure by using a different approximation to the functional KL-divergence. Recent work has also proposed to adapt BNN priors to interpretable functional priors by minimising the Wasserstein distance between a BNN prior and a Gaussian process [Tran et al., 2020]. Another line of research exploits the Wasserstein gradient flow and tries to encourage diversity in the function space [D'Angelo et al., 2021, D'Angelo and Fortuin, 2021].

Gaussian processes Standard Gaussian process regression [Rasmussen and Williams, 2006] allows interpretable prior specification but scales poorly with respect to the number of data points. As a result, a plethora of approximation techniques are introduced. On one hand, there are variational approximations to the true posterior [Titsias, 2009a, Hensman et al., 2013] and several extensions [Hensman et al., 2017, Salimbeni et al., 2018, Dutordoir et al., 2020]. On the other hand, GPU utilisation is combined with Krylov subspace methods to obtain scalability [Gardner et al., 2018, Wang et al., 2019].

4.3 Background

In this section we give some background on generalised variational inference in infinite dimensions and introduce Gaussian measures in Hilbert spaces. We further discuss their relation to the more familiar Gaussian processes at the end.

4.3.1 Generalised Variational Inference in Function Spaces

In functional variational inference, we assign a prior $p(f)$ to the unknown function $f \in E$, where E is a function space[2]. The prior is combined with the likelihood $p(y|f)$ to give the posterior $p(f|y)$. The posterior is often intractable which is why in variational inference we specify a tractable variational

[2]We assume E to be a Polish space, which avoids technical difficulties in defining the posterior measure [Ghosal and van der Vaart, 2017, Chapter 1.3]

approximation $q(f)$ to $p(f|y)$ and train our model by maximising the evidence lower bound (ELBO)

$$\mathcal{L} = \mathbb{E}_{q(f)}\big[\log p(y|f)\big] - \mathbb{D}_{\mathrm{KL}}\big(q(f), p(f)\big), \tag{4.1}$$

where $\mathbb{D}_{\mathrm{KL}}$ denotes the KL divergence. Note that in the case where E is infinite dimensional $p(f)$ and $q(f)$ cannot be probability density functions with respect to the Lebesgue measure [see e.g. Hunt et al., 1992, for a discussion], which is why the above notation, although commonly used, is imprecise. What we in fact mean are the probability measures over E associated with the prior and variational approximation. We will denote these measures as $\mathbb{P}^F$ and $\mathbb{Q}^F$ from now on to make this difference explicit. The ELBO in this notation reads as

$$\mathcal{L} := \mathbb{E}_{\mathbb{Q}}\big[\log p(y|F)\big] - \mathbb{D}_{\mathrm{KL}}\big(\mathbb{Q}^F, \mathbb{P}^F\big). \tag{4.2}$$

Note that the KL divergence (for measures) is defined as

$$\mathbb{D}_{\mathrm{KL}}\big(\mathbb{Q}^F, \mathbb{P}^F\big) = \int \log\left(\frac{d\mathbb{Q}^F}{d\mathbb{P}^F}(f)\right) d\mathbb{Q}^F(f), \tag{4.3}$$

where we assume that $\mathbb{Q}^F$ is dominated by the measure $\mathbb{P}^F$ which guarantees the existence of the Radon-Nikodym derivative $d\mathbb{Q}^F/d\mathbb{P}^F$. A number of papers focus on obtaining tractable approximations of (4.3) [Sun et al., 2018, Rudner et al., 2020, Ma and Hernández-Lobato, 2021]. However, the use of KL-divergence in infinite-dimensional function spaces can be a delicate task, since benign constructions of priors and variational approximations may not satisfy that $\mathbb{Q}^F$ is dominated by $\mathbb{P}^F$ which leads to $\mathbb{D}_{\mathrm{KL}}\big(\mathbb{Q}^F, \mathbb{P}^F\big) = \infty$ [Burt et al., 2020a]. This often renders the objective (4.2) useless or at least problematic.

A *true Bayesian* is committed to the use of the KL divergence in (4.2) as maximising $\mathcal{L}$ is equivalent to minimising the KL divergence between the true posterior measure and the variational measure. This equivalence is typically demonstrated using pdfs but the argument generalises to infinite dimensions as is shown for GPs in Matthews et al. [2016] or in a more measure theoretic formulation in Theorem 4 of Wild and Wynne [2022].

However, Knoblauch et al. [2019] argue that given the problems of prior and likelihood specification as well as available compute, an axiomatically justified way of moving from prior to posterior beliefs is by solving a more general optimisation problem [Knoblauch et al., 2019, Theorem 15]. Crucially it is valid to replace the KL-divergence by an arbitrary measure of dissimilarity $\mathbb{D}$ satisfying $\mathbb{D}(\mathbb{Q}^F, \mathbb{P}^F) \geq 0$ and $\mathbb{D}(\mathbb{Q}^F, \mathbb{P}^F) = 0 \Rightarrow \mathbb{Q}^F = \mathbb{P}^F$. The arguments in Knoblauch et al. [2019] are made assuming the existence

of a pdf for the prior, but they rely solely on a reformulation of Bayesian inference as optimization problem [Knoblauch et al., 2019, Chapter 2]. We show in Appendix B.1 that this reformulation can also be made for infinite-dimensional prior measures and therefore consider the generalized loss

$$\mathcal{L} := -\mathbb{E}_{\mathbb{Q}}\big[\log p(y|F)\big] + \mathbb{D}\big(\mathbb{Q}^F, \mathbb{P}^F\big), \tag{4.4}$$

a valid optimization objective for an arbitrary dissimilarity measure $\mathbb{D}$. This is merely an infinite-dimensional version of equation (10) in Knoblauch et al. [2019]. We refer to inference targeting the objective (4.4) as *generalised variational inference in function space* (GVI-FS) [3]. Generalised variational inference can be interpreted as regularised loss minimisation lifted into the space of probability measures. The first term in (4.4) is understood as a loss which we want to minimise on average, while the second term punishes strong deviations from the prior.

The particular instance of GVI-FS that we explore is where both $\mathbb{P}^F$ and $\mathbb{Q}^F$ are Gaussian measures (on an infinite-dimensional Hilbert space) and $\mathbb{D}$ is chosen to be the Wasserstein metric [Kantorovich, 1960]. We will refer to this setting as *Gaussian Wasserstein Inference in Function Space* (GWI-FS) or more concisely as *Gaussian Wasserstein Inference* (GWI)

4.3.2 Gaussian Random Elements and Gaussian Measures in Hilbert Spaces

In this section we introduce Gaussian random elements (GRE) and Gaussian measures in Hilbert spaces – these concepts are somewhat technical but crucial in the construction of our method. We then describe their close relationship to the more familiar Gaussian process notions in the next section.

Let $\big(\Omega, \mathcal{A}, \mathbb{P}\big)$ be the underlying (physical) probability space and $\big(H, \langle\cdot,\cdot\rangle\big)$ be a Hilbert space.

Gaussian random elements A measurable function $F : \Omega \to H$ is called GRE (in H) if and only if $\langle F, h\rangle : \Omega \to \mathbb{R}$ has a scalar Gaussian distribution for all $h \in H$.[4] Every GRE F has a mean element $m \in H$ defined by

$$m := \int F(\omega)\, d\mathbb{P}(\omega) \tag{4.5}$$

and a (linear) covariance operator $C : H \to H$ defined by

$$Ch(\cdot) := \int \langle F(\omega), h\rangle F(\omega)\, \mathbb{P}(\omega) - \langle m, h\rangle m. \tag{4.6}$$

[3]The term GVI typically refers to a situation where the negative log-likelihood is replaced by an arbitrary loss function, and the KL-divergence is replaced by an arbitrary dissimilarity measure. In this paper, we are primarily focused on comparing with GP based methods, so we aimed to make minimal changes and, as a result, kept the negative log-likelihood unchanged. However, the theory and method carry over mutatis mutandis.

[4]We allow for the degenerate case where the variance of $\langle F, h\rangle$ is zero. This means we interpret a Gaussian with variance zero as Dirac measure.

for $h \in H$. Both integrals are to be understood as Bochner integrals [Kukush, 2020, Chapter 3]. The Bochner integral has the property that $\left\langle \int F(\omega)\, d\mathbb{P}(\omega), h \right\rangle = \int \langle F(\omega), h\rangle \, d\mathbb{P}(\omega)$ for all $h \in H$. This combined with Fubini's theorem and the definition of a GRE implies that

$$\langle F, h\rangle \sim \mathcal{N}\big(\langle m, h\rangle, \langle Ch, h\rangle\big), \tag{4.7}$$

for any $h \in H$ with $\mathcal{N}(\mu, \sigma^2)$ denoting the normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Similarly we denote $F \sim \mathcal{N}(m, C)$ for a GRE in H with mean element m and covariance operator C. It can be shown that the covariance operator C of a GRE is a positive self-adjoint trace class operator. Conversely, for every positive self-adjoint trace class operator and every $m \in H$, there exists a GRE with $F \sim \mathcal{N}(m, C)$ [Bogachev, 1998, Theorem 2.3.1].

Gaussian measures The push-forward measure of $\mathbb{P}$ through F is defined as $\mathbb{P}^F(A) := \mathbb{P}\big(F^{-1}(A)\big)$ for all Borel-measurable $A \subset H$. If $F \sim \mathcal{N}(m, C)$ is a GRE, we call $P := \mathbb{P}^F$ a GM and write $P = \mathcal{N}(m, C)$. Note that GMs or equivalently GREs allow us to specify probability distributions over (infinite-dimensional) Hilbert spaces by using a given mean element and a given covariance operator.

Details about Gaussian Measures in Hilbert spaces can be found in Chapter 2 of Da Prato and Zabczyk [2014] or in Kukush [2020]. In fact, Gaussian measures can be defined on even more general linear spaces such as Banach or Fréchet spaces [Bogachev, 1998].

4.3.3 Gaussian Processes and Their Corresponding Measures

In this section we describe how Gaussian processes – a standard tool to assign functional priors in Bayesian machine learning – are related to Gaussian measures.

Let $\big(\Omega, \mathcal{A}, \mathbb{P}\big)$ be the underlying (physical) probability space and $\mathcal{X} \subset \mathbb{R}^D$ be measurable. The (product-) measurable mapping $G : \Omega \times \mathcal{X} \to \mathbb{R}$ is called a Gaussian process (GP) if and only if for all $N \in \mathbb{N}$ and all $X = \{x_n\}_{n=1}^N \subset \mathcal{X}$ the random vector $G(X) := \big(G(\cdot, x_1), \ldots, G(\cdot, x_N)\big)^T$ is multivariate Gaussian. For a GP G we define a mean function $m(x) := \mathbb{E}\big[G(x)\big]$, $x \in \mathcal{X}$, and a covariance function by $k(x, x') := \mathbb{C}[G(x), G(x')\big]$ for $x, x' \in \mathcal{X}$. Here $\mathbb{E}$ denotes the expected value and $\mathbb{C}[\cdot, \cdot]$ the covariance. It follows from the definition that $G(X) \sim \mathcal{N}\big(m(X), k(X, X)\big)$ for any $\{x_n\}_{n=1}^N \subset \mathcal{X}$, where we define $m(X) := \big(m(x_n)\big)_{n=1}^N$ and $k(X, X) := \big(k(x_n, x_{n'})\big)_{n,n'=1}^N$. We write $G \sim GP(m, k)$ for a GP with mean function m and covariance function k. Note that by the properties of the covariance we know that $k(X, X)$ is a (symmetric) positive semi-definite matrix for all $\{x_n\}_{n=1}^N \subset \mathcal{X}$ and $N \in \mathbb{N}$. A function with this property is called *kernel*, a terminology that we adopt henceforth. Kolmogorov's existence theorem [Billingsley, 2008, Section 36] guarantees the existence of a Gaussian process for any kernel k and any

mean function m. The standard reference for Gaussian processes in machine learning is Rasmussen and Williams [2006].

The main advantage of Gaussian processes in specifying priors over a function space is that the kernel k allows us to incorporate readily interpretable prior assumptions, such as smoothness or periodicity. For example, choosing the squared exponential kernel [Rasmussen and Williams, 2006] implies that the unknown function is infinitely differentiable and that the correlation of the functional output is higher the closer the inputs are.

In order to insert the Gaussian process prior into our generalised loss in (4.4) we need to know the probability measure that is associated to the Gaussian process. In general, we can associate more than one Gaussian measure with a given Gaussian process. For example:

- If the GP has continuous sample paths we can associate a Gaussian measure on the space E of continuous functions with it [Lifshits, 2012, Example 2.4].
- If the GP has square-integrable sample paths we can associate a Gaussian measure on the Hilbert space of square-integrable functions with it (cf. Theorem 4.3.1).

These sample path properties can be guaranteed under additional assumptions on the kernel. The next theorem discusses one such kernel condition which guarantees the GP to have sample paths in the Hilbert space of square integrable functions, denoted $L^2(\mathcal{X}, \rho, \mathbb{R})$, with inner product $\langle g, h \rangle_2 := \int_{\mathcal{X}} g(x)h(x)\, d\rho(x)$.

Theorem 4.3.1. *Let $F \sim GP(m, k)$ be a GP with mean $m \in L^2(\mathcal{X}, \rho, \mathbb{R})$ and kernel k such that*

$$\int_{\mathcal{X}} k(x, x)\, d\rho(x) < \infty. \tag{4.8}$$

We call a kernel satisfying (4.8) trace-class kernel. Then the mapping $\widetilde{F} : \Omega \to L^2(\mathcal{X}, \rho, \mathbb{R})$ defined as $\widetilde{F}(\omega) := F(\omega, \cdot)$ is a Gaussian random element with mean m and covariance operator C given as

$$Cg(\cdot) := \int k(\cdot, x')g(x')\, d\rho(x') \tag{4.9}$$

for any $g \in L^2(\mathcal{X}, \rho, \mathbb{R})$. Consequently $P := \mathbb{P}^F \sim \mathcal{N}(m, C)$ is a Gaussian measure.

Proof. The fact that $\widetilde{F}$ as defined above is a GRE follows immediately from Example 2.3.16 in Bogachev [1998]. The fact that m is its mean and C as defined in (4.9) is its covariance operator follows from Fubini's theorem. □

It shall be noted that there is no need to appeal to GPs in order to justify the use of GMs. In fact, it has

recently been demonstrated that variational inference for GPs can be formulated purely in terms of GMs [Wild and Wynne, 2022]. In the following sections we will therefore deploy GMs without any reference to GPs, but it is of course always possible to think of them as the measures that correspond to GPs where the kernel satisfies an additional assumption such as (4.8).

4.4 Gaussian Wasserstein Inference in Function Spaces

This section describes how the Wasserstein distance between Gaussian measures can be used to obtain a tractable optimization target for inference in function spaces. In the end, we discuss several parameterisations of GWI and introduce our main inference method - the GWI-net.

4.4.1 Model Description

Let $\{(x_n, y_n)\}_{n=1}^N \subset \mathcal{X} \times \mathcal{Y}$ be $N \in \mathbb{N}$ paired observations. We assume that $\mathcal{X} \subset \mathbb{R}^D$, $D \in \mathbb{N}$ and further that $\mathcal{Y} = \mathbb{R}$ for regression and $\mathcal{Y} = \{1, \dots, J\}$ for classification with $J \in \mathbb{N}$ classes. We focus in our exposition here on the regression case but have given the relevant derivations for classification in Appendix B.6.

As pointed out in section 4.3.1, GVI in function space minimises the generalized loss $\mathcal{L} = -\mathbb{E}_{\mathbb{Q}}\big[\log p(y|F)\big] + \mathbb{D}\big(\mathbb{Q}^F, \mathbb{P}^F\big)$. We make the mild assumption that the unknown function f is square integrable with respect to the data distribution ρ on $\mathcal{X}$ which means $f \in E = L^2(\mathcal{X}, \rho, \mathbb{R})$. The prior $P := \mathbb{P}^F$ is described by a Gaussian measure with mean $m_P \in \mathcal{L}^2(\mathcal{X}, \rho, \mathbb{R})$ and covariance operator C_P described by a trace-class kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which means it is given as $(C_P f)(x) := \int_{\mathcal{X}} k(x, x') f(x') \, d\rho(x')$ for all $f \in L^2(\mathcal{X}, \rho, \mathbb{R})$. We assume a Gaussian likelihood for $y := (y_1, \dots, y_N)$ given as $p(y|f) := \prod_{n=1}^N p(y_n|f)$[5] with

$$p(y_n|f) := \mathcal{N}(y_n \,|\, f(x_n), \sigma^2), \tag{4.10}$$

where $\mathcal{N}(\cdot \,|\, \mu, \sigma^2)$ denotes the pdf of a normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. This prior and likelihood are natural choices as they mimic the standard formulation of Gaussian process regression. The variational approximation of the posterior is chosen to be another Gaussian measure $Q := \mathbb{Q}^F$ with arbitrary mean $m_Q \in L^2(\mathcal{X}, \rho, \mathbb{R})$ and arbitrary covariance operator C_Q induced by a trace-class kernel r: $(C_Q f)(x) := \int_{\mathcal{X}} r(x, x') f(x') \, d\rho(x')$ for all $f \in L^2(\mathcal{X}, \rho, \mathbb{R})$.

It remains for us to select a dissimilarity measure $\mathbb{D}$. As already pointed out in the introduction we decide to use the Wasserstein distance W_2 (a formal definition is given in Appendix B.3). This choice was guided

[5] Astute readers may notice that the definition of the likelihood contains a pointwise evaluation $f(x_n)$ which may not be a well defined operation on $L^2(\mathcal{X}, \rho, \mathbb{R})$. We detail in Appendix B.12 how that problem can be circumvented and that in fact $F(x) \sim \mathcal{N}(m(x), k(x, x))$ as one would expected.

by two considerations:

1. The Wasserstein metric was proven to be a useful metric for probability distributions in machine learning applications [Arjovsky et al., 2017, Tran et al., 2020].

2. The Wasserstein distance is tractable for arbitrary Gaussian measures on (separable) Hilbert spaces [Gelbrich, 1990] and given as

$$W_2^2(P,Q) = ||m_P - m_Q||_2^2 + tr(C_P) + tr(C_Q) - 2 \cdot tr\Big[\big(C_P^{1/2} C_Q C_P^{1/2}\big)^{1/2}\Big], \tag{4.11}$$

   where tr denotes the trace of an operator and $C_P^{1/2}$ is the square root of the positive, self-adjoint operator C_P. This is in stark contrast to the KL-divergence that is infinite whenever $\mathbb{Q}^F$ is not dominated by $\mathbb{P}^F$ and even in the case where it is finite there exists no explicit formula for the KL-divergence in infinite dimensions.

The generalized loss for our model is therefore given as

$$\mathcal{L} = -\sum_{n=1}^{N} \mathbb{E}_{\mathbb{Q}}\Big[\log \mathcal{N}\big(y_n \,|\, F(x_n), \sigma^2\big)\Big] + W_2(P,Q). \tag{4.12}$$

Note that the expected log-likelihood in (4.12) can be calculated analytically as

$$\mathbb{E}_{\mathbb{Q}}\Big[\log \mathcal{N}\big(y_n \,|\, F(x_n), \sigma^2\big)\Big] = -\frac{N}{2}\log(2\pi\sigma^2) - \sum_{n=1}^{N} \frac{\big(y_n - m_Q(x_n)\big)^2 + r(x_n,x_n)}{2\sigma^2}. \tag{4.13}$$

It remains to produce an approximation of (4.11) in order to obtain a tractable inference procedure. To this end, note that by definition $||m_P - m_Q||_2^2 = \int \big(m_P(x) - m_Q(x)\big)^2 \, d\rho(x)$ and further $tr(C_P) = \int k(x,x)\, d\rho(x)$ [Brislawn, 1991]. We now replace the true input distribution ρ with the empirical data distribution $\widehat{\rho} := \frac{1}{N}\sum_{n=1}^{N} \delta_{x_n}$, where δ_x denotes the Dirac measure in $x \in \mathcal{X}$. This gives $||m_P - m_Q||_2^2 \approx \frac{1}{N}\sum_{n=1}^{N} \big(m_P(x_n) - m_Q(x_n)\big)^2$, $tr(C_P) \approx \frac{1}{N}\sum_{n=1}^{N} k(x_n,x_n)$ and $tr(C_Q) \approx \frac{1}{N}\sum_{n=1}^{N} r(x_n,x_n)$. It remains to provide an approximation of $tr\Big[\big(C_P^{1/2} C_Q C_P^{1/2}\big)^{1/2}\Big]$. The key idea is to approximate the spectrum of $C_P^{1/2} C_Q C_P^{1/2}$ by that of an appropriate kernel matrix. Details are discussed in Appendix B.4. This leads to the following final approximation for the Wasserstein metric

$$\hat{W}^2 := \frac{1}{N}\sum_{n=1}^{N} \big(m_P(x_n) - m_Q(x_n)\big)^2 + \frac{1}{N}\sum_{n=1}^{N} k(x_n,x_n) \tag{4.14}$$

$$+ \frac{1}{N}\sum_{n=1}^{N} r(x_n,x_n) - \frac{2}{\sqrt{N N_S}} \sum_{s=1}^{N_S} \sqrt{\lambda_s\big(r(X_S,X)k(X,X_S)\big)}, \tag{4.15}$$

where $X_S := (x_{S,1}, \dots, x_{S,N_S})$ with $x_{S,1}, \dots x_{S,N_S} \in \mathbb{R}^D$ being subsampled from the input data X. Further $r(X_S, X) := \big(r(x_{S,s}, x_n)\big)_{s,n}$ and $k(X, X_S) := \big(k(x_n, x_{S,s})\big)_{n,s}$ for $n = 1, \dots, N$, $s = 1, \dots, N_S$ and $\lambda_s\big(r(X_S, X)k(X, X_S)\big)$ denotes the s-th eigenvalue of the matrix $r(X_S, X)k(X, X_S) \in \mathbb{R}^{N_S \times N_S}$. The approximation quality of $\widehat{W}$ is related to the spectral decay of the operator $C_P C_Q$, which in turn is determined by the kernels k and r. For the choices made in Section 4.4.2 we empirically observe rapid spectral decay (cp. Appendix B.13) and therefore are confident that the 2-Wasserstein distance is estimated reliably for our method.

The combination of (4.13), (4.14) and (4.15) gives a generalised loss that is tractable in terms of m_P, m_Q, k, and r. If we disregard computation time of m_P, m_Q, k and r, the generalized loss can be evaluated in $\mathcal{O}(N + N_S^2 N + N_S^3)$, where typically $N_S \ll N$, e.g. $N_S = 100$. We provide a batch version of our loss in Appendix B.5 which reduces the computations to $\mathcal{O}(N_S^2 N_B + N_S^3)$ where $N_B \ll N$ is the batch-size. Note, however, that the final computation time for our method will be determined by the complexity hidden in the evaluation of m_Q, m_P, k, and r as we need N_B evaluations of m_Q and m_P and $N_S \cdot N_B$ evaluations of r and k per iteration.

4.4.2 Parameterisations of Prior and Variational Measure

The prior for our model is given as $P = \mathcal{N}(m_P, C_P)$ with C_P induced by a trace-class kernel k. One of the advantages of the proposed approach is that any trace-class kernel is allowed and this is where one can incorporate specific assumptions and domain expertise. This is a thoroughly studied topic: the prior kernel can encode periodicity [Durrande et al., 2016], geometric intuition [van der Wilk et al., 2018], and even model linear constraints for the unknown function [Jidling et al., 2017]. In order to keep the exposition simple and maintain focus on the inference, however, and in line with using simple priors on network weights in standard Bayesian deep learning, we opt for a simple zero mean prior $m_P = 0$ and a standard ARD kernel k given as

$$k(x, x') = \sigma_f^2 \exp\Big(-\frac{1}{2} \sum_{d=1}^{D} \frac{(x_d - x'_d)^2}{\alpha_d^2} \Big) \tag{4.16}$$

for $x, x' \in \mathcal{X} \subset \mathbb{R}^D$. We refer to $\sigma_f > 0$ as *kernel scaling factor* and to $\alpha_d > 0$ as *length-scale* for dimension d. The parameters σ_f and $\alpha := (\alpha_1, \dots, \alpha_D)$ are called *prior hyperparameters*.

The rest of the section explores various choices for the variational mean m_Q and the variational kernel r. The parameters appearing in the specification of m_Q and r are referred to as *variational parameters*.

GWI: Stochastic variational Gaussian process Let $z_1, \dots, z_M \in \mathcal{X}$ be a subsample of the data X

with $M \ll N$. We define the posterior mean

$$m_Q(x) := m_P(x) + \sum_{m=1}^{M} \beta_m k_m(x) \tag{4.17}$$

with $\beta_m \in \mathbb{R}$ and $k_m(x) := k(x, z_m)$, $m = 1, \ldots, M$ where k is the prior kernel k and $\beta := (\beta_1, \ldots, \beta_M) \in \mathbb{R}^M$ are variational parameters. Define further the variational kernel

$$r(x, x') = k(x, x') - k_Z(x)^T k(Z, Z)^{-1} k_Z(x) + k_Z(x)^T \Sigma k_Z(x), \tag{4.18}$$

where $\Sigma \in \mathbb{R}^{M \times M}$ is the symmetric and positive definite variational covariance matrix that parameterises r. This choice of m_Q and r essentially recovers the *stochastic variational Gaussian processes* (SVGP) model of Titsias [2009a]. Note that in our framework it is straightforward to use all (or just more) basis functions for the mean $m_Q(x) := m_P(x) + \sum_{n=1}^{N} \beta_n k_n(x)$ where $k_n(x) := k(x, x_n)$, $\beta_n \in \mathbb{R}$, $n = 1, \ldots, N$. This mirrors the construction in Cheng and Boots [2017] where we allow more parameters to learn the mean than in SVGP. However, both Titsias [2009a] and Cheng and Boots [2017] use a different objective function than GWI to learn the unknown parameters.

GWI: deep neural network with SVGP An interesting approach is to parameterise the posterior mean as a deep neural network (DNN). We assume the DNN has $L \in \mathbb{N}$ hidden layers and the width of layer $\ell = 1, \ldots, L$ is denoted D_ℓ with $D_0 := D$ and $D_{L+1} = 1$. This means we define $g^1(x) := W^1 x + b^1$ and further $h^\ell(x) := \phi\big(g^\ell(x)\big)$, $g^{\ell+1}(x) := W^{\ell+1} h^\ell(x) + b^{\ell+1}$ for $\ell = 1, \ldots, L$. Here W^{l+1} is $D_{\ell+1} \times D_\ell$ matrix, $b^{\ell+1} \in \mathbb{R}^{D_{\ell+1}}$ is a bias vector for layer l and ϕ an activation function. We can then define the variational mean as $m_Q(x) := m_P(x) + g^{L+1}(x)$. If we choose the SVGP kernel r in (4.18), we essentially predict with a neural network and quantify uncertainty with a (sparse) Gaussian process, capturing the beneficial properties of both.

Neural networks have been combined in several ways with GPs [Wilson et al., 2016, Tran et al., 2020]. However, to the best of our knowledge they were not used to directly parameterise the posterior in the context of generalized variational inference in function space. The spirit of our approach is fundamentally different: rather than thinking of a neural network as a model which needs to be made Bayesian, we use it as a parametrisation of a variational posterior.

Note that we do not provide an exhaustive study of how to best parameterise the variational measure, since this paper is focused on demonstrating the ability of the proposed method to obtain valid uncertainty quantification. An exploratory study on how properties and quality of uncertainty quantification relate to different choices of m_Q and r is reserved for future work. We mention potential problems that can occur

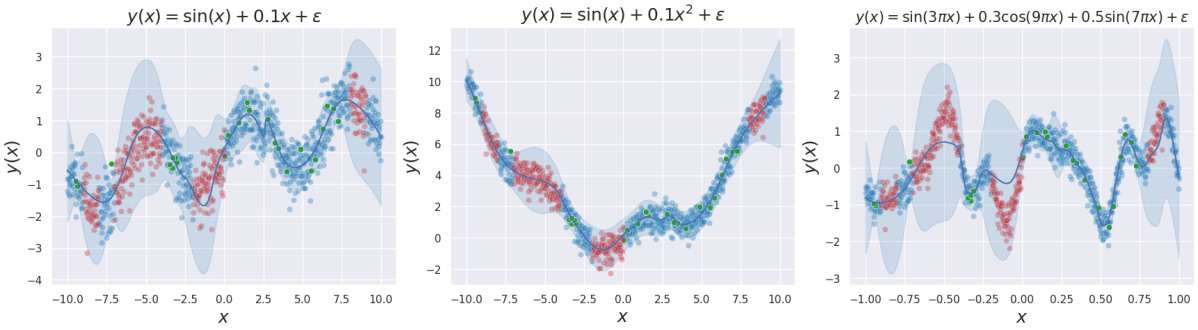


Figure 4.1: : Training data : Unseen data : Inducing points
We query the above functions at $N = 1000$ equidistant points and add white noise with $\epsilon \sim \mathcal{N}(0, 0.5^2)$. We use $M = 30$ inducing points and train our method as described in Appendix B.7. The plot shows $m_Q(x) \pm 1.96\sqrt{\mathbb{V}[Y^*(x)|Y]}$ where $\mathbb{V}[Y^*(x)|Y]$ is the posterior predictive variance given as $r(x,x)+\sigma^2$.

from misspecification in Appendix B.10.

4.5 Experiments

We show results for GWI with the SVGP mean (4.17) and the SVGP kernel (4.18). We use the shorthand GWI: SVGP for this approach. Additionally we implement the DNN mean with the SVGP kernel (4.18). This combination achieves impressive results on various regression and classification tasks. We call this method GWI: DNN-SVGP or simply GWI-net.

Illustrative Examples In Figure 4.1 we illustrate GWI-net on a few toy examples. One can clearly see that the posterior predictive variance expands for regions lacking observations which demonstrates the ability of our method to quantify uncertainty. We provide an additional graphic comparison with SVGP in Appendix B.12 and an example for two-dimensional inputs in Appendix B.9

There we show that the pathologies regarding the quantification of in-between uncertainty discussed in Foong et al. [2020] are not present for our method.

UCI Regression In Table 4.1 we report the average test negative log-likelihood (NLL) (cf. Appendix B.7 for details) of GWI: SVGP and GWI-net (GWI: DNN-SVGP) and the results of several weight-space approaches for BNNs: Bayes-by-Backprop (BBB) [Blundell et al., 2015], variational dropout (VDO) [Gal and Ghahramani, 2016], and variational alpha dropout ($\alpha = 0.5$) [Li and Gal, 2017]. We also compare with four function-space BNN inference methods: functional variational inference with BNN prior (FVI) [Ma and Hernández-Lobato, 2021], variationally implicit processes (VIP) with BNNs, VIP-Neural processes [Ma et al., 2019], and functional BNNs (FBNNs) [Sun et al., 2018]. In order to ensure a fair comparison we matched neural network architectures and training procedures for the different methods. Detailed explanations are given in Appendix B.7.

Dataset	N	D	GWI		FVI	VIP-BNN	VIP-NP	BBB	VDO	$\alpha = 0.5$	FBNN	**EXACT GP**
			SVGP	DNN-SVGP								
BOSTON	506	13	2.8±0.31	**2.27±0.06**	2.33±0.04	2.45±0.04	2.45±0.03	2.76±0.04	2.63±0.10	2.45±0.02	2.30±0.10	2.46±0.04
CONCRETE	1030	8	3.24±0.09	**2.64±0.06**	2.88±0.06	3.02±0.02	3.13±0.02	3.28±0.01	3.23±0.01	3.06±0.03	3.09±0.01	3.05±0.02
ENERGY	768	8	1.81±0.19	0.91±0.12	0.58±0.05	**0.56±0.04**	0.60±0.03	2.17±0.02	1.13±0.02	0.95±0.09	0.68±0.02	0.54±0.02
KIN8NM	8192	8	-0.86±0.38	**-1.2±0.03**	-1.15±0.01	-1.12±0.01	-1.05±0.00	-0.81±0.01	-0.83±0.01	-0.92±0.02	N/A±0.00	N/A±0.00
POWER	9568	4	3.35±0.22	2.74±0.02	**2.69±0.00**	2.92±0.00	2.90±0.00	2.83±0.01	2.88±0.00	2.81±0.00	N/A±0.00	N/A±0.00
PROTEIN	45730	9	**2.84±0.04**	2.87±0.0	2.85±0.00	2.87±0.00	2.96±0.02	3.00±0.00	2.99±0.00	2.90±0.00	N/A±0.00	N/A±0.00
RED WINE	1588	11	0.97±0.02	**0.76±0.08**	0.97±0.06	0.97±0.02	1.20±0.04	1.01±0.02	0.97±0.02	1.01±0.02	1.04±0.01	0.26±0.03
YACHT	308	6	2.37±0.55	0.29±0.1	0.59±0.11	**-0.02±0.07**	0.59±0.13	1.11±0.04	1.22±0.18	0.79±0.11	1.03±0.03	0.10±0.05
NAVAL	11934	16	**-7.25±0.08**	-6.76±0.1	-7.21±0.06	-5.62±0.04	-4.11±0.00	-2.80±0.00	-2.80±0.00	-2.97±0.14	-7.13±0.02	N/A±0.00
Mean Rank			5.5	**2.06**	2.22	3.33	4.94	7	6.11	4.83		

Table 4.1: The table shows the average test NLL on several UCI regression datasets. We train on random 90% of the data and predict on 10%. This is repeated 10 times and we report mean and standard deviation. The results for our competitors are taken from Ma and Hernández-Lobato [2021].

One can see that GWI-net obtains the best mean rank of all methods being the best model on 4/9 datasets and performing competitively on all datasets. Note that we exclude FBNN and exact Gaussian processes from the comparison because their computational complexity is often prohibitively large.

Classification and OOD Detection We demonstrate the ability of GWI to perform image classifications on Fashion MNIST [Xiao et al., 2017] and CIFAR-10 [Krizhevsky et al., 2009]. We compare to FVI, mean-field variational inference (MVFI) [Blundell et al., 2015], maximum a posteriori approximation (MAP), K-FAC Laplace-GNN [Martens and Grosse, 2015] and its dampened version [Ritter et al., 2018]. Implementation details are discussed in B.8.

We also assess the ability of our model to perform out-of-distribution detection using in-distribution (ID) / out of-distribution (OOD) pairs given as FashionMNIST/MNIST and CIFAR10/SVNH. Following the setting of Osawa et al. [2019], Immer et al. [2021] we calculate the area under the curve (AUC) of a binary out-of-distribution classifier based on predictive entropies. Results are shown in Table 4.2.

	FMNIST			CIFAR 10		
Model	Accuracy	NLL	OOD-AUC	Accuracy	NLL	OOD-AUC
GWI-net	**93.25 ±0.09**	**0.250 ±0.00**	**0.959 ±0.01**	**83.82 ±0.00**	**0.553 ±0.00**	0.618 ±0.00
FVI	91.60±0.14	0.254±0.05	0.956±0.06	77.69 ±0.64	0.675±0.03	0.883±0.04
MFVI	91.20±0.10	0.343±0.01	0.782±0.02	76.40±0.52	1.372±0.02	0.589±0.01
MAP	91.39±0.11	0.258±0.00	0.864±0.00	77.41±0.06	0.690±0.00	0.809±0.01
KFAC-LAPLACE	84.42±0.12	0.942±0.01	0.945±0.00	72.49±0.20	1.274±0.01	0.548±0.01
RITTER et al.	91.20±0.07	0.265±0.00	0.947±0.00	77.38±0.06	0.661±0.00	0.796±0.00

Table 4.2: We report average accuracy, NLL and OOD-AUC on test data for 10 different train/test splits. The results for FVI are obtained from Ma and Hernández-Lobato [2021] and for MAP, KFAC and Ritter et al. results are taken from Immer et al. [2021] .

Our method performs best in all categories on the Fashion MNIST dataset achieving state-of-the-art results. On CIFAR10 we obtain the highest accuracy and best NLL by a significant margin and perform competitively in the OOD detection task.

4.6 Limitations

In this section we discuss some of the shortcomings and difficulties which are related to our method.

The GVI-FS framework allows the specification of function space inference via infinite dimensional parameters such as mean and kernel functions. This great flexibility essentially allows the specification of mismatched prior and posterior parameters. We illustrate such a case in Appendix B.10.

GWI-net relies on the SVGP kernel defined in 4.18 for its posterior approximation. It therefore inherits numerical instabilities associated with the inversion of the kernel matrix. For the data sets discussed in this paper it was possible to overcome these issues by smart initialisation of the optimiser (cf. Appendix B.7), but it may be an interesting research avenue to come up with a kernel that avoids these instabilities.

Our method approximates the Wasserstein distance in function space via the spectrum of kernel matrices (cf. Appendix B.4). These approximations require quick spectral decay of the composition of prior and variational covariance operator to be accurate and computationally tractable. The prior SE kernel combined with the variational SVGP kernel did have this property (cf. B.13) which allowed for cheap and accurate approximations. However, other parameterisations may result in less accurate estimation. A theoretical investigation of how the approximation quality relates to kernel properties is an interesting topic for further research.

The proposed framework models prior and variational distribution with a Gaussian measure on the space of square integrable functions. As a consequence the posterior distribution for the functional output is Gaussian as well. This means it is unimodal and concentrated around the posterior mean. Although this constrains the form of functional posterior significantly the authors would argue that the empirical success of GWI-net demonstrates that the approach is flexible to meaningfully quantify uncertainty.

4.7 Conclusion

In this paper, we developed a framework for generalised variational inference in infinite-dimensional function spaces. We leveraged the function space perspective to develop a new inference approach combining Gaussian measures and Wasserstein distance with predictive performance of deep neural networks, yielding principled uncertainty quantification. The value of our method was demonstrated on several benchmark datasets.

Statement of Authorship for joint/multi-authored papers for PGR thesis

Title of Paper	Generalised Variational Inference in Function Spaces: Gaussian Measures meet Bayesian Deep Learning
Publication Status	published
Publication Details	Veit D. Wild*, Robert Hu* and Dino Sejdinovic. “Generalised Variational Inference in Function Spaces: Gaussian Measures meet Bayesian Deep Learning.” Advances in Neural Information Processing Systems (NeurIPS), 2022.

Student Confirmation

Student Name:	Veit David Wild		
Contribution to the Paper	• Derivation of GVI in function space, • Derivation of GWI algorithm, • derivation of all theoretical results in the paper, • writing of the manuscript.		
Signature	V.Wild	Date	28.01.2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Dino Sejdinovic		
Supervisor comments			
Signature	Dino Sejdinovic	Date	30 January 2024

This completed form should be included in the thesis, at the end of the relevant chapter.

5 | A Rigorous Link between Deep Ensembles and (Variational) Bayesian Methods

This chapter is based on the following publication:

> **Veit D. Wild**, Sahra Ghalebikesabi, Dino Sejdinovic and Jeremias Knoblauch. "A Rigorous Link between Deep Ensembles and (Variational) Bayesian Methods". *Advances in Neural Information Processing Systems (NeurIPS), 2023.*

Abstract

We establish the first mathematically rigorous link between Bayesian, variational Bayesian, and ensemble methods. A key step towards this is to reformulate the non-convex optimisation problem typically encountered in deep learning as a convex optimisation in the space of probability measures. On a technical level, our contribution amounts to studying generalised variational inference through the lens of Wasserstein gradient flows. The result is a unified theory of various seemingly disconnected approaches that are commonly used for uncertainty quantification in deep learning—including deep ensembles and (variational) Bayesian methods. This offers a fresh perspective on the reasons behind the success of deep ensembles over procedures based on standard variational inference, and allows the derivation of new ensembling schemes with convergence guarantees. We showcase this by proposing a family of interacting deep ensembles with direct parallels to the interactions of particle systems in thermodynamics, and use our theory to prove the convergence of these algorithms to a well-defined global minimiser on the space of probability measures.

5.1 Introduction

A major challenge in modern deep learning is the accurate quantification of uncertainty. To develop trustworthy AI systems, it will be crucial for them to recognize their own limitations and to convey the inherent uncertainty in the predictions. Many different approaches have been suggested for this. In variational inference (VI), a prior distribution for the weights and biases in the neural network is assigned and the best approximation to the Bayes posterior is selected from a class of parameterised distributions [Graves, 2011, Blundell et al., 2015, Gal and Ghahramani, 2016, Louizos and Welling, 2017]. An alternative approach is to (approximately) generate samples from the Bayes posterior via Monte Carlo methods [Welling and Teh, 2011, Neal, 2012]. Another approach, known as deep ensembles, relies on a train-and-repeat heuristic to quantify uncertainty [Lakshminarayanan et al., 2017].

Much ink has been spilled over whether one can see deep ensembles as a Bayesian procedure [Wilson, 2020, Izmailov et al., 2021, D'Angelo and Fortuin, 2021] and over how these seemingly different methods might relate to each other. Building on this discussion, we shed further light on the connections between Bayesian inference and deep ensemble techniques by taking a different vantage point. In particular, we show that methods as different as variational inference, Langevin sampling [Ermak, 1975], and deep ensembles can be derived from a well-studied generally infinite-dimensional regularised optimisation problem over the space of probability measures [see e.g. Guedj and Shawe-Taylor, 2019, Knoblauch et al., 2019]. As a result, we find that the differences between these algorithms map directly onto different choices regarding this optimisation problem. Key differences between the algorithms boil down to different choices of regularisers, and whether they implement a finite-dimensional or infinite-dimensional gradient descent. Here, finite-dimensional gradient descent corresponds to parameterised VI schemes, whilst the infinite-dimensional case maps onto ensemble methods.

The contribution of this paper is a new theory that generates insights into existing algorithms and provides links between them that are mathematically rigorous, unexpected, and useful. On a technical level, our innovation consists in analysing them as algorithms that target an optimisation problem in the space of probability measures through the use of a powerful technical device: the Wasserstein gradient flow [see e.g. Ambrosio et al., 2005]. While the theory is this paper's main concern, its potentially substantial methodological payoff is demonstrated through the derivation of a new inference algorithm based on gradient descent in infinite dimensions and regularisation with the maximum mean discrepancy. We use our theory to show that this algorithm—unlike standard deep ensembles—is derived from a strictly convex objective defined over the space of probability measures. Thus, it targets a unique minimum, and is capable of producing samples from this global minimiser in the infinite particle and time horizon limit.

This makes the algorithm provably convergent; and we hope that it can help plant the seeds for renewed innovations in theory-inspired algorithms for (Bayesian) deep learning.

The paper proceeds as follows: Section 5.2 discusses the advantages of lifting losses defined on Euclidean spaces into the space of probability measures through a generalised variational objective. Section 5.3 introduces the notion of Wasserstein gradient flows, while Section 5.4 links them to the aforementioned objective and explains how they can be used to construct algorithms that bridge Bayesian and ensemble methods. The paper concludes with Section 5.5, where the findings of the paper are illustrated numerically.

5.2 Convexification through Probabilistic Lifting

One of the most technically challenging aspects of contemporary machine learning theory is that the losses $\ell : \mathbb{R}^J \to \mathbb{R}$ we wish to minimise are often highly non-convex. For instance, one could wish to minimise $\ell(\theta) := \frac{1}{N}\sum_{n=1}^{N}\big(y_n - f_\theta(x_n)\big)^2$ where $\{(x_n, y_n)\}_{n=1}^N$ is a set of paired observations and f_θ a neural network with parameters θ. While deep learning has shown that non-convexity is often a negligible *practical* concern, it makes it near-impossible to prove many basic *theoretical* results that a good learning theory is concerned with, as ℓ has many local (or global) minima [Fort et al., 2019, Wilson and Izmailov, 2020]. We reintroduce convexity by lifting the problem onto a computationally more challenging space. In this sense, the price we pay for the convenience of convexity is the transformation of a finite-dimensional problem into an infinite-dimensional one, which is numerically more difficult to tackle. Figure 5.1 illustrates our approach:

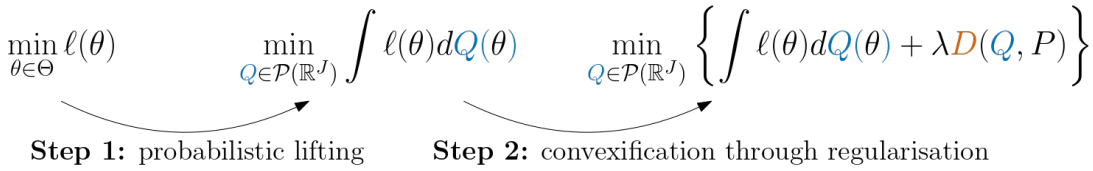


Figure 5.1: Illustration of convexification through probabilistic lifting.

First, we transform a non-convex optimisation $\min_{\theta\in\Theta} \ell(\theta)$ into an infinite-dimensional optimisation over the set of probability measures $\mathcal{P}(\mathbb{R}^J)$, yielding $\min_{Q\in\mathcal{P}(\mathbb{R}^J)} \int \ell(\theta) dQ(\theta)$. As an integral, this objective is linear in Q. However, linear functions are not strictly convex. We therefore need to add a strictly convex regulariser to ensure uniqueness of the minimiser.[1] We prove in Appendix C.1 that indeed—for a regulariser $D : \mathcal{P}(\mathbb{R}^J) \times \mathcal{P}(\mathbb{R}^J) \to [0, \infty]$ such that $(Q, P) \mapsto D(Q, P)$ is strictly convex and $P \in \mathcal{P}(\mathbb{R}^J)$ a fixed measure—existence and uniqueness of a global minimiser can be guaranteed.

[1] To illustrate why this is necessary, assume that there are θ_A and θ_B in $\mathbb{R}^J$ so that $\ell(\theta_A) = \ell(\theta_B) = \min_{\theta\in\Theta} \ell(\theta)$. Due to linearity, each measure $Q_t \in \mathcal{P}(\mathbb{R}^J)$, $t \in [0, 1]$, defined as $Q_t := (1 - t)\delta_{\theta_A} + t\delta_{\theta_B}$ provides one (of the infinitely many) global minima in the set $\arg\min_{Q\in\mathcal{P}(\mathbb{R}^J)} \int \ell(\theta)\, dQ(\theta)$.

Given a scaling constant $\lambda > 0$, we can now put everything together to obtain the loss L and the unique minimiser Q^* as

$$L(Q) := \int \ell(\theta)\, dQ(\theta) + \lambda D(Q, P), \qquad Q^* := \operatorname*{arg\,min}_{Q \in \mathcal{P}(\mathbb{R}^J)} L(Q). \tag{5.1}$$

Throughout, whenever Q and Q^* have an associated Lebesgue density, we write them as q and q^*. Moreover, all measures, densities, and integrals will be defined on the parameter space $\mathbb{R}^J$ of θ. Similarly, the gradient operator ∇ will exclusively denote differentiation with respect to $\theta \in \mathbb{R}^J$.

5.2.1 One Objective with Many Interpretations

In the current paper, our sole focus lies on resolving the difficulties associated with non-convex optimisation of ℓ on Euclidean spaces. Through probabilistic lifting and convexification, we can identify a unique minimiser Q^* in the new space, which minimises the θ-averaged loss $\theta \mapsto \ell(\theta)$ without deviating too drastically from some reference measure P. In this sense, Q^* summarises the quality of all (local and global) minimisers of $\ell(\theta)$ by assigning them a corresponding weight. The choices for λ, D and P determine the trade-off between the initial loss ℓ and reference measure P and therefore the weights we assign to different solutions.

Yet, (5.1) is not a new problem form: it has various interpretations, depending on the choices for D, λ, ℓ and the framework of analysis [see e.g. Knoblauch et al., 2019, for a discussion]. For example, if $\ell(\theta)$ is a negative log likelihood, D is the Kullback–Leibler divergence (KL), and $\lambda = 1$, Q^* is the **standard Bayesian** posterior, and P is the Bayesian prior. This interpretation of P as a prior carries over to **generalised Bayesian** methods, in which we can choose $\ell(\theta)$ to be any loss, D to be any divergence on $\mathcal{P}(\mathbb{R}^J)$, and λ to regulate how fast we learn from data [see e.g. Bissiri et al., 2016, Jewson et al., 2018, Knoblauch et al., 2018, Miller and Dunson, 2019, Knoblauch, 2019, Alquier, 2021a, Husain and Knoblauch, 2022, Matsubara et al., 2022, Wild et al., 2022, Wu and Martin, 2023, Altamirano et al., 2023]. In essence, the core justification for these generalisations is that the very assumptions justifying application of Bayes' Rule are violated in modern machine learning. In practical terms, this results in a view of Bayes' posteriors as one—of many possible—measure-valued estimators Q^* of the form in (5.1). Once this vantage point is taken, it is not clear why one *should* be limited to using only one particular type of loss and regulariser for *every* possible problem. Seeking a parallel with optimisation on Euclidean domains, one may then compare the orthodox Bayesian view with the insistence on only using quadratic regularisation for *any* problem. While it is beyond the scope of this paper to cover these arguments in depth, we refer the interested reader to Knoblauch et al. [2019].

A second line of research featuring objectives as in (5.1) are **PAC-Bayes** methods, whose aim is to construct generalisation bounds [see e.g. Shawe-Taylor and Williamson, 1997, McAllester, 1999a,9, Grünwald, 2011]. Here, ℓ is a general loss, but P only has the interpretation of some reference measure that helps us measure the complexity of our hypotheses via $Q \mapsto \lambda D(Q, P)$ [Guedj and Shawe-Taylor, 2019, Alquier, 2021b]. Classic PAC-Bayesian bounds set D to be KL, but there has been a recent push for different complexity measures [Alquier and Guedj, 2018, Bégin et al., 2016, Haddouche and Guedj, 2023].

5.2.2 Generalised Variational Inference (GVI) in Finite and Infinite Dimensions

In line with the terminology coined in Knoblauch et al. [2019], we refer to any algorithm aimed at solving (5.1) as a **generalised variational inference (GVI)** method. Broadly speaking, there are two ways one could design such algorithms: in finite or infinite dimensions.

Finite-dimensional GVI: This is the original approach advocated for in Knoblauch et al. [2019]: instead of trying to compute Q^*, approximate it by solving $Q_{\nu^*} = \arg\min_{Q_\nu \in \mathcal{Q}} L(Q)$ for a set of measures $\mathcal{Q} := \{Q_\nu : \nu \in \Gamma\} \subset \mathcal{P}(\mathbb{R}^J)$ parameterised by a parameter $\nu \in \Gamma \subseteq \mathbb{R}^I$. To find $\mathcal{Q}_\nu$, one now simply performs (finite-dimensional) gradient descent with respect to the function $\nu \mapsto L(Q_\nu)$. For the special case where P is a Bayesian prior, $\lambda = 1$, $\ell(\theta)$ is a negative log likelihood parametrised by θ, and $D = \mathrm{KL}$, this recovers the well-known standard VI algorithm. To the best of our knowledge, all methods that refer to themselves as VI or GVI in the context of deep learning are based on this approach [see e.g. Graves, 2011, Blundell et al., 2015, Louizos and Welling, 2017, Wild et al., 2022]. Since procedures of this type solve a finite-dimensional version of (5.1), we refer to them as **finite-dimensional GVI (FD-GVI)** methods throughout the paper. While such algorithms can perform well, they have some obvious theoretical problems: First of all, the finite-dimensional approach typically forces us to choose $\mathcal{Q}$ and P to be simple distributions such as Gaussians to ensure that $L(Q_\nu)$ is a tractable function of ν. This often results in a $\mathcal{Q}$ that is unlikely to contain a good approximation to Q^*; raising doubt if Q_{ν^*} can approximate Q^* in any meaningful sense. Secondly, even if $Q \mapsto L(Q)$ is strictly convex on $\mathcal{P}(\mathbb{R}^J)$, the parameterised objective $\nu \in \Gamma \mapsto L(Q_\nu)$ is usually not. Hence, there is no guarantee that gradient descent leads us to Q_{ν^*}. This point also applies to very expressive variational families [Rezende and Mohamed, 2015, Mescheder et al., 2017] which may be sufficiently rich that $Q^* \in \mathcal{Q}$, but whose optimisation problem $\nu \in \Gamma \mapsto L(Q_\nu)$ is typically non-convex and hard to solve, so that no guarantee for finding Q^* can be provided. While this does not necessarily make FD-GVI impractical, it does make it exceedingly difficult to provide a rigorous theoretical analysis outside of narrowly defined settings.

FD-GVI in function space: An collection of approaches formulated as infinite-dimensional problems

are GVI methods on an infinite-dimensional function space [Ma et al., 2019, Sun et al., 2018, Ma and Hernández-Lobato, 2021, Rodriguez-Santana et al., 2022, Wild et al., 2022]. Here, the loss is often convex in function space. In practice however, the variational stochastic process still requires parameterization to be computationally feasible—and in this sense, function space methods are FD-GVI approaches. The resulting objectives require a good approximation of the functional KL-divergence (which is often challenging), and lead to a typically highly non-convex variational optimization problem in the parameterised space.

Infinite-dimensional GVI: Instead of minimising the (non-convex) problem $\nu \mapsto L(Q_\nu)$, we want to exploit the convex structure of $Q \mapsto L(Q)$. Of course, a priori it is not even clear how to compute the gradient for a function $Q \mapsto L(Q)$ defined on an infinite-dimensional nonlinear space such as $\mathcal{P}(\mathbb{R}^J)$. However, in the next part of this paper we will discuss that it is possible to implement a gradient descent in infinite dimensions by using **gradient flows** on a metric space of probability measures [Ambrosio et al., 2005]. More specifically, one can solve the optimisation problem (5.1) by following the curve of steepest descent in the 2-Wasserstein space. As it turns out, this approach is not only theoretically sound, but also conceptually elegant: it unifies existing algorithms for uncertainty quantification in deep learning, and even allows us to derive new ones. We refer to algorithms based on some form of infinite-dimensional gradient descent as **infinite-dimensional GVI (ID-GVI)**. Infinite-dimensional gradient descent methods have recently gained attention in the machine learning community. For existing methods of this kind, the goal is to generate samples from a target $\widehat{Q}$ that has a *known* form (such as the Bayes posterior) by applying a gradient flow to $Q \in \mathcal{P}(\mathbb{R}^J) \mapsto E(Q, \widehat{Q})$ where $E(\cdot, \cdot)$ is a discrepancy measure. Some methods apply the Wasserstein gradient flow (WGF) for different choices of E [Arbel et al., 2019, Korba et al., 2021, Glaser et al., 2021], whilst other methods like Stein variational gradient descent (SVGD) [Liu and Wang, 2016] stay within the Bayesian paradigm ($E = \mathrm{KL}$, $\widehat{Q}$ = Bayes posterior) but use a gradient flow other than the WGF [Liu, 2017]. D'Angelo and Fortuin [2021] exploit the WGF in the standard Bayesian context and combine it with different gradient estimators [Li and Turner, 2017, Shi et al., 2018] to obtain repulsive deep ensembling schemes. Note that this is different from our methods as our repulsive effect is induced by the regulariser and not a consequence of the gradient-estimators. Since our focus is on tackling the problems associated with non-convex optimisation in Euclidean space, the approach we propose is inherently different from all of these existing methods: our target Q^* is only implicitly defined via (5.1), and not known explicitly.

5.3 Gradient Flows in Finite and Infinite Dimensions

Before we can realise our ambition to solve (5.1) with an ID-GVI scheme, we need to cover the relevant bases. To this end, we will discuss gradient flows in finite and infinite dimensions, and explain how they can be used to construct infinite-dimensional gradient descent schemes. In essence, a gradient flow is the limit of a gradient descent whose step size goes to zero. While the current section introduces this idea for the finite-dimensional case for ease of exposition, its use in constructing algorithms within the current paper will be for the infinite-dimensional case.

Gradient descent finds local minima of losses $\ell : \mathbb{R}^J \to \mathbb{R}$ by iteratively improving an initial guess $\theta_0 \in \mathbb{R}^J$ through the update

$$\theta_{k+1} := \theta_k - \eta \nabla \ell(\theta_k),\ k \in \mathbb{N},$$

where $\eta > 0$ is a step-size and $\nabla \ell$ denotes the gradient of ℓ. For sufficiently small $\eta > 0$, this update can equivalently be written as

$$\theta_{k+1} = \underset{\theta \in \mathbb{R}^J}{\arg\min} \left\{ \ell(\theta) + \frac{1}{2\eta} ||\theta - \theta_k||_2^2 \right\}. \tag{5.2}$$

Gradient flows formalise the following logic: for any fixed η, we can continuously interpolate the corresponding gradient descent iterates $\{\theta_k\}_{k \in \mathbb{N}_0}$. To do this, we simply define a function $\theta^\eta : [0, \infty) \to \mathbb{R}$ as $\theta^\eta(t) := \theta_{t/\eta}$ for $t \in \eta\mathbb{N}_0 := \{0, \eta, 2\eta, ...\}$. For $t \notin \eta\mathbb{N}_0$ we linearly interpolate[2]. As $\eta \to 0$, the function θ^η converges to a differentiable function $\theta_* : [0, \infty) \to \mathbb{R}$ called the gradient flow of ℓ, because it is characterised as solution to the ordinary differential equation (ODE) $\theta_*'(t) = -\nabla \ell(\theta_*(t))$ with initial condition $\theta_*(0) = \theta_0$. Intuitively, $\theta_*(t)$ is a continuous-time version of discrete-time gradient descent; and navigates through the loss landscape so that at time t, an infinitesimally small step in the direction of steepest descent is taken. Put differently: gradient descent is nothing but an Euler discretisation of the gradient flow ODE [see also Santambrogio, 2017]. The result is that for mathematical convenience, one often analyses discrete-time gradient descent as though it were a continuous gradient flow—with the hope that for sufficiently small η, the behaviour of both will essentially be the same.

Our results in the infinite-dimensional case follow this principle: we propose an algorithm based on discretisation, but use continuous gradient flows to guide the analysis. To this end, the next section generalises gradient flows to the nonlinear infinite-dimensional setting.

[2]This means $\theta^\eta(t) := \frac{\theta_{s+1} - \theta_s}{\eta}\big(t - s\eta\big) + \theta_s$, for $t \in [s\eta, (s+1)\eta)$ and $s \in \mathbb{N}_0$.

5.3.1 Gradient Flows in Wasserstein Spaces

Let $\mathcal{P}_2(\mathbb{R}^J)$ be the space of probability measures with finite second moment equipped with the 2-Wasserstein metric given as

$$W_2(P,Q)^2 = \inf\left\{\int ||\theta - \theta'||_2^2 \, d\pi(\theta,\theta') : \pi \in \mathcal{C}(P,Q)\right\} \tag{5.3}$$

where $\mathcal{C}(P,Q) \subset \mathcal{P}(\mathbb{R}^J \times \mathbb{R}^J)$ denotes the set of all probability measure on $\mathbb{R}^J \times \mathbb{R}^J$ such that $\pi(A \times \mathbb{R}^J) = P(A)$ and $\pi(\mathbb{R}^J \times B) = Q(B)$ for all $A, B \subset \mathbb{R}^J$ [see also Chapter 6 of Villani et al., 2009]. Further, let $L : \mathcal{P}_2(\mathbb{R}^J) \to (-\infty, \infty]$ be some functional—for example L in (5.1). In direct analogy to (5.2), we can improve upon an initial guess $Q_0 \in \mathcal{P}_2(\mathbb{R}^J)$ by iteratively solving

$$Q_{k+1} := \underset{Q \in \mathcal{P}_2(\mathbb{R}^J)}{\arg\min} \left\{L(Q) + \frac{1}{2\eta} W_2(Q, Q_k)^2\right\} \tag{5.4}$$

for $k \in \mathbb{N}_0$ and small $\eta > 0$ [see Chapter 2 of Ambrosio et al., 2005, for details]. Again, for $\eta \to 0$, an appropriate limit yields a continuously indexed family of measures $\{Q(t)\}_{t \geq 0}$. If L is sufficiently smooth and $Q_0 = Q(0)$ has Lebesgue density q_0, the time evolution for the corresponding pdfs $\{q(t)\}_{t \geq 0}$ is given by the partial differential equation (PDE)

$$\partial_t q(t,\theta) = \nabla \cdot \Big(q(t,\theta) \, \nabla_W L\big[Q(t)\big](\theta)\Big), \tag{5.5}$$

with $q(0,\cdot) = q_0$ [Villani, 2003, Section 9.1]. Here $\nabla \cdot f := \sum_{j=1}^J \partial_j f_j$ denotes the divergence operator and $\nabla_W L[Q] : \mathbb{R}^J \to \mathbb{R}^J$ the Wasserstein gradient (WG) of L at Q. For the purpose of this paper, it is sufficient to think of the WG as a gradient of the first variation; i.e. $\nabla_W L[Q] = \nabla L'[Q]$ where $L'[Q] : \mathbb{R}^J \to \mathbb{R}^J$ is the first variation of L at Q [Villani et al., 2009, Exercise 15.10]. The Wasserstein gradient flow (WGF) for L is then the solution $q^\dagger$ to the PDE (5.5). If L is chosen as in (5.1), our hope is that the logic of finite-dimensional gradient descent carries over; and that $\lim_{t \to \infty} q^\dagger(t,\cdot)$ is in fact the density q^* corresponding to Q^*.

Following this reasoning, this paper applies the WGF for (5.1) to obtain an ID-GVI algorithm. In Section 5.4 and Appendices C–F, we formally show that the WGF indeed yields Q^* (in the limit as $t \to \infty$) for a number of regularisers of practical interest.

5.3.2 Realising the Wasserstein Gradient Flow

In theory, the PDE in (5.5) could be solved numerically in order to implement the infinite-dimensional gradient descent for (5.1). In practice however, this is impossible: numerical solutions to PDEs become

computationally infeasible for the high-dimensional parameter spaces which are common in deep learning applications. Rather than trying to first approximate the q solving (5.5) and then sampling from its limit in a second step, we will instead formulate equations which replicate how the samples from the solution to (5.5) evolve in time. This leads to tractable inference algorithms that can be implemented in high dimensions.

Given the goal of producing samples directly, we focus on a particular form of loss that is well-studied in the context of thermodynamics [Santambrogio, 2015, Chapter 7], and which recovers various forms of the GVI problem in (5.1) (see Section 5.4). In thermodynamics, $Q \in \mathcal{P}_2(\mathbb{R}^J)$ describes the distribution of particles located at specific points in $\mathbb{R}^J$. The overall energy of a collection of particles sampled from Q is decomposed into three parts: (i) the external potential $V(\theta)$ which acts on each particle individually, (ii) the interaction energy $\kappa(\theta, \theta')$ describing pairwise interactions between particles, and (iii) an overall entropy of the system measuring how concentrated the distribution Q is. Taking these components together, we obtain the so called **free energy**

$$L^{\text{fe}}(Q) := \int V(\theta)\, dQ(\theta) + \frac{\lambda_1}{2} \iint \kappa(\theta, \theta')\, dQ(\theta) dQ(\theta') + \lambda_2 \int \log q(\theta) q(\theta)\, d\theta, \tag{5.6}$$

for $Q \in \mathcal{P}_2(\mathbb{R}^J)$ with Lebesgue density q, $\lambda_1 \geq 0$, $\lambda_2 \geq 0$. Note that for $\lambda_2 > 0$ we implicitly assume that Q has a density. Following Section 9.1 in Villani et al. [2009], its WG is

$$\nabla_W L^{\text{fe}}[Q](\theta) = \nabla V(\theta) + \lambda_1 \int (\nabla_1 \kappa)(\theta, \theta')\, dQ(\theta') + \lambda_2 \nabla \log q(\theta),$$

where $\theta \in \mathbb{R}^J$, and $\nabla_1 \kappa$ denotes the gradient of κ with respect to the first variable. We plug this into (5.5) to obtain the desired density evolution. Importantly, this time evolution has the exact form of a nonlinear Fokker-Planck equation associated with a stochastic process of McKean-Vlasov type (see Appendix C.2 for details). Fortunately for us, it is well-known that such processes can be approximated through interacting particles [Veretennikov, 2006] generated by the following procedure:

Step 1: Sample $N_E \in \mathbb{N}$ particles $\theta_1(0), \ldots, \theta_{N_E}(0)$ independently from $Q_0 \in \mathcal{P}_2(\mathbb{R}^J)$.

Step 2: Evolve the particle θ_n by following the stochastic differential equation (SDE)

$$d\theta_n(t) = -\Big(\nabla V\big(\theta_n(t)\big) + \frac{\lambda_1}{N_E} \sum_{j=1}^{N_E} (\nabla_1 \kappa)\big(\theta_n(t), \theta_j(t)\big)\Big) dt + \sqrt{2\lambda_2} dB_n(t), \tag{5.7}$$

for $n = 1, \ldots, N_E$, and $\{B_n(t)\}_{t>0}$ stochastically independent Brownian motions.

As $N_E \to \infty$, the distribution of $\theta_1(t), \ldots, \theta_{N_E}(t)$ evolves in t in the same way as the sequence of

densities $q(t, \cdot)$ solving (5.5). This means that we can implement infinite-dimensional gradient descent by following the WGF and simulating trajectories for infinitely many interacting particles according to the above procedure. In practice, we can only simulate finitely many trajectories over a finite time horizon. This produces samples $\theta_1(T), \ldots, \theta_{N_E}(T)$ for $N_E \in \mathbb{N}$ and $T > 0$. Our intuition and Section 5.4 tell us that, as desired, the distribution of $\theta_1(T), \ldots, \theta_{N_E}(T)$ will be close to the global minimiser of L^{fe}.

In the next section, we will use the above algorithm to construct an ID-GVI method producing samples approximately distributed according to Q^* defined in (5.1). Since N_E and T are finite, and since we need to discretise (C.123), there will be an approximation error. Given this, how good are the samples produced by such methods? As we shall demonstrate in Section 5.5, the approximation errors are small, and certainly should be expected to be much smaller than those of standard VI and other FD-GVI methods.

5.4 Optimisation in the Space of Probability Measures

With the WGF on thermodynamic objectives in place, we can now finally show how it yields ID-GVI algorithms to solve (5.1). We put particular focus on the analysis of the regulariser D; providing new perspectives on heuristics for uncertainty quantification in deep learning in the process. Specifically, we establish formal links explaining how they may (not) be understood as a Bayesian procedure. Beyond that, we derive the WGF associated with regularisation using the maximum mean discrepancy, and provide a theoretical analysis of its convergence properties.

5.4.1 Unregularised Probabilistic Lifting: Deep Ensembles

We start the analysis with the base case of an unregularised functional $L(Q) = \int \ell(\theta) dQ(\theta)$, corresponding to $\lambda = 0$ in (5.1). This is also a special case of (5.6) with $\lambda_1 = \lambda_2 = 0$. As $\lambda_1 = 0$, there is no interaction term, and all particles can be simulated independently from one another as

$$\theta_1(0), \ldots, \theta_{N_E}(0) \sim Q_0, \quad \theta_n'(t) = -\nabla \ell\big(\theta_n(t)\big), \quad n = 1, \ldots, N_E.$$

This simple algorithm happens to coincide exactly with how deep ensembles (DEs) are constructed [see e.g. Lakshminarayanan et al., 2017]. In other words: the simple heuristic of running gradient descent algorithm several times with random initialisations sampled from Q_0 is an approximation of the WGF for the *unregularised* probabilistic lifting of the loss function ℓ.

Following the WGF in this case does not generally produce samples from a global minimiser of L. Indeed, the fact that L generally does not even have a unique global minimiser was the motivation for regularisation in (5.1). Even if L had a unique minimiser however, a DE would not find it. The result

below proves this formally: unsurprisingly, deep ensembles simply sample the local minima of ℓ with a probability that depends on the domain of attraction and the initialisation distribution Q_0.

Theorem 5.4.1. *If ℓ has countably many local minima $\{m_i : i \in \mathbb{N}\}$, then it holds independently for each $n = 1, \dots, N_E$ that*

$$\theta_n(t) \xrightarrow{\mathcal{D}} \sum_{i=1}^{\infty} Q_0(\Theta_i)\, \delta_{m_i} =: Q_\infty$$

for $t \to \infty$. Here $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution and $\Theta_i = \{\theta \in \mathbb{R}^J : \lim_{t\to\infty} \theta_*(t) = m_i$ and $\theta_*(0) = \theta\}$ denotes the domain of attraction for m_i with respect to the gradient flow θ_*.

A proof with technical assumptions—most importantly a version of the famous Lojasiewicz inequality—is in Appendix C.3. Theorem 5.4.1 derives the limiting distribution for $\theta_1(T), \dots, \theta_{N_E}(T)$, which shows that—unless all local minima are global minima—the WGF does not generate samples from a global minimum of $Q \mapsto L(Q)$ for the unregularised case $\lambda = 0$. Note that his result is not directly applicable for the over-parameterised situation encountered in deep learning as the set of minimisers is not countable here [Liu et al., 2022]. However, we chose to include Theorem 5.4.1 as it illustrates the dependency of Q_∞ on the initialisation Q_0. This will remain true for typical deep learning losses but in this case there is no way of compactly writing Q_∞.

However, despite these theoretical shortcomings, DEs remain highly competitive in practice and typically beat FD-GVI methods like standard VI [Ovadia et al., 2019, Fort et al., 2019]. This is perhaps not surprising: DEs implement an infinite-dimensional gradient descent, while FD-GVI methods are parametrically constrained. Perhaps more surprisingly, we observe in Section 5.5 that DEs can even easily compete with the more theoretically sound and regularised ID-GVI methods that will be discussed in Section 5.4.2 and 5.4.3. We study this phenomenon in Section 5.5, and find that it is a consequence of the fact that in deep learning, N_E is small compared to the number of local minima (cf. Figure 5.4).

5.4.2 Regularisation with the Kullback–Leibler Divergence: Deep Langevin Ensembles

In Section 5.2, we argued for regularisation by D to ensure a unique minimiser Q^*. The Kullback–Leibler divergence ($D = \mathrm{KL}$) is the canonical choice for (generalised) Bayesian and PAC-Bayesian methods [Bissiri et al., 2016, Knoblauch et al., 2019, Guedj and Shawe-Taylor, 2019, Alquier, 2021b]. Now, Q^* has a known form: if P has a pdf p, it has an associated density given by $q^*(\theta) \propto \exp(-\frac{1}{\lambda}\ell(\theta))p(\theta)$ [Knoblauch et al., 2019, Theorem 1].

Notice that the KL-regularised version of L in (5.1) can be rewritten in terms of the objective L^{fe} in (5.6) by setting $V(\theta) = \ell(\theta) - \lambda \log p(\theta)$, $\lambda_1 = 0$ and $\lambda_2 = \lambda$. Compared to the unregularised objective of the previous section (where $\lambda = 0$), the external potential is now adjusted by $-\lambda \log p(\theta)$, forcing Q^* to

allocate more mass in regions where p has high density. Beyond that, the presence of the negative entropy term has three effects: it ensures that the objective is strictly convex, that Q^* is more spread out, and that it has a density q^*. Since $\lambda_1 = 0$, the corresponding particle method still does not have an interaction and is given as

$$\theta_1(0), \dots, \theta_{N_E}(0) \sim Q_0 \quad d\theta_n(t) = -\nabla V\big(\theta_n(t)\big)dt + \sqrt{2\lambda}dB_n(t), \quad n = 1, \dots, N_E. \tag{5.8}$$

Clearly, this is just the Langevin SDE and we call this approach the **deep Langevin ensemble (DLE)**. While the name may suggest that DLE is equivalent to the unadjusted Langevin algorithm (ULA) [Roberts and Tweedie, 1996], this is not so: for T discretisation steps $t_1, t_2, \dots t_T$, DLE approximates measures using the *end-points* of N_E trajectories given by $\{\theta_n(t_T)\}_{n=1}^{N_E}$. In contrast, ULA would use a (sub)set of the samples $\{\theta_1(t_i)\}_{i=1}^{T}$ generated from one single particle's *trajectory*. To analyse DLEs, we build on the Langevin dynamics literature: in Appendix C.4, we show that $\text{Law}[\theta_n(t)] \xrightarrow{\mathcal{D}} Q^*$ as $t \to \infty$, independently for each $n = 1, \dots, N_E$. Hence $\theta_1(T), \dots, \theta_{N_E}(T)$ will for large $T > 0$ be approximately distributed according to Q^*. Comparing DE and DLE in this light, we note several important key differences: Q^* as defined per (5.1) is unique, has the form of a Gibbs measure, and can be sampled from using (5.8). In contrast, unregularised DE produces samples from Q_∞ in Theorem 5.4.1 which is not the global minimiser. Specifically neither Q_∞ nor Q^* for DEs correspond to the Bayes posterior. It is therefore not a Bayesian procedure in any commonly accepted sense of the word.

5.4.3 Regularisation with Maximum Mean Discrepancy: Deep Repulsive Langevin Ensembles

Regularising with KL is attractive because Q^* has a known form. However, in our theory, there is no reason to restrict attention to a single type of regulariser: we introduced D to convexify our objective. It is therefore of theoretical and practical interest to see which algorithmic effects are induced by other regularisers. We illustrate this by first considering regularisation using the squared maximum-mean discrepancy (MMD) [see e.g. Gretton et al., 2012] only, and then a combination of MMD and KL.

For a kernel $\kappa : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}$, the squared MMD between measures Q and P is

$$\begin{aligned}\text{MMD}(Q, P)^2 = &\iint \kappa(\theta, \theta')dQ(\theta)dQ(\theta') - 2\iint \kappa(\theta, \theta')dQ(\theta)dP(\theta') \\ &+ \iint \kappa(\theta, \theta')dP(\theta)dP(\theta').\end{aligned}$$

MMD measures the difference between within-sample similarity and across-sample similarity, so it is smaller when samples from P are similar to samples from Q, but also larger when samples within Q

are similar to each other. This means that regularising (5.1) with $D = \mathrm{MMD}^2$ introduces interactions characterised precisely by the kernel κ, and we can show this explicitly by rewriting L of (5.1) into the form of L^{fe} in (5.6). In other words, inclusion of MMD^2 makes particles repel each other, making it more likely that they fall into different (rather than the same) local minima. Writing the kernel mean embedding as $\mu_P(\theta) := \int \kappa(\theta, \theta')\, dP(\theta')$, we see that up to a constant not depending on Q, $L(Q) = L^{\mathrm{fe}}(Q)$ for $V(\theta) = \ell(\theta) - \lambda_1 \mu_P(\theta)$, $\lambda = \frac{\lambda_1}{2}$, and $\lambda_2 = 0$. While we can show that a global minimiser Q^* exists, and while we could produce particles using the algorithm of Section 5.3.2, we cannot guarantee that they are distributed according to Q^* (see Appendix C.6). Essentially, this is because in certain situations, we cannot guarantee that Q^* has a density for $D = \mathrm{MMD}^2$.

To remedy this problem, we additionally regularise with the KL: since $\mathrm{KL}(Q, P) = \infty$ if P has a Lebesgue density but Q has not, this now guarantees that Q^* has a density q^*. In terms of (5.1), this means that $D = \lambda\, \mathrm{MMD}^2 + \lambda'\, \mathrm{KL}$. Adding regularisers like this has a long tradition, and is usually done to combine the different strengths of various regularisers [see e.g. Zou and Hastie, 2005]. Here, we follow this logic: the KL ensures that Q^* has a density, and the MMD makes particles repel each other. With this, we can rewrite $L(Q)$ in terms of $L^{\mathrm{fe}}(Q)$ up to a constant not depending on Q by taking $\lambda = \frac{\lambda_1}{2}$, $\lambda' = \lambda_2$, and $V(\theta) = \ell(\theta) - \lambda_1 \mu_P(\theta) - \lambda_2 \log p(\theta)$. Using the same algorithmic blueprint as before, we evolve particles according to (5.7). As these particles follow an augmented Langevin SDE that incorporates repulsive particle interactions via κ, we call this method the **deep repulsive Langevin ensemble (DRLE)**. We show in Theorem (5.4.2) (cf. Appendix C.5 for details) that DRLEs generate samples from the global minimiser Q^* in the infinite particle and infinite time horizon limit.

Theorem 5.4.2. *Let $Q^{n,N_E}(t)$ be the distribution of $\theta_n(t)$, $n = 1, \ldots, N_E$, generated via (5.7). Then*

$$\lim_{t\to\infty} \lim_{N_E\to\infty} Q^{n,N_E}(t) = Q^* \text{ (in distribution)}$$

for each $n = 1, \ldots, N_E$ whenever the corresponding the McKean-Vlasov process in (5.7) converges to a unique invariant measure.

This is remarkable: we have constructed an algorithm that generates samples from the global minimiser Q^*—even though a formal expression for what exactly Q^* looks like is unknown! This demonstrates how impressively powerful the WGF is as tool to derive inference algorithms. Note that this is completely different from sampling methods employed for Bayesian methods, for which the form of Q^* is typically known explicitly up to a proportionality constant.

A notable shortcoming of Theorem 5.4.2 is its asymptotic nature. A more refined analysis would quantify how fast the convergence happens in terms of N_E, T, the SDE's discretisation error, and potentially even

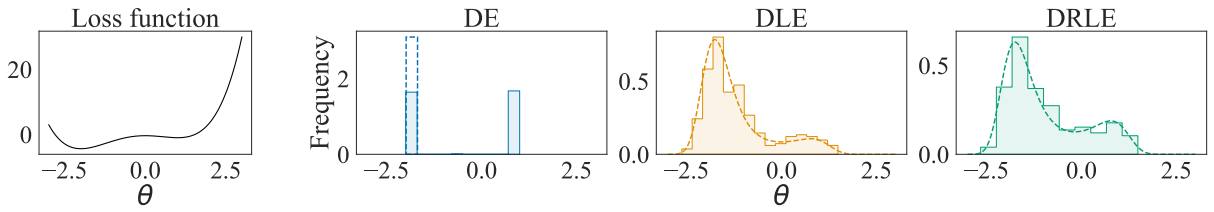


Figure 5.2: We generate $N_E = 300$ particles from DE, DLE and DRLE. The theoretically optimal global minimisers Q^* are depicted with dashed line strokes, and the generated samples are displayed via histograms. We use $P = Q_0 = \mathcal{N}(0, 1)$ for DLE and DRLE. Notice that the optimal Q^* differs slightly between DLE and DRLE.

the use of unbiased estimators for the loss based on sub-sampling. While the existing literature could be adapted to derive the speed of convergence for DRLE in T [Ambrosio et al., 2005, Section 11.2], this would require a strong convexity assumption on the potential V, which will not be satisfied for any applications in deep learning. This is perhaps unsurprising: even for the Langevin algorithm—probably the most thoroughly analysed algorithm in this literature—no convergence rates have been derived that are applicable to the highly multi-modal target measures encountered in Bayesian deep learning [Wibisono, 2019, Chewi et al., 2022].

That being said, for the case of deep learning, FD-GVI approaches never come with any convergence guarantees. Indeed, in this setting they even fail to provide basic asymptotic guarantees. As a consequence, the fact that it is even possible for us to provide *any* asymptotic guarantees using realistic assumptions marks an improvement over the available theory for FD-GVI methods, and (by virtue of Theorem 5.4.1) over DEs as well.

5.5 Experiments

Since the paper's primary focus is on theory, we use two experiments to reinforce some of the predictions it makes in previous sections, and a third experiment that shows why–in direct contradiction to a naive interpretation of the presented theory–it is typically difficult to beat simple DEs. More details about the conducted experiments can be found in Appendix C.7. The code is available on `https://anonymous.4open.science/r/GVI-WGF-5963`.

Global minimisers: Figure 5.2 illustrates the theory of Sections 5.4 and Appendices C–F: DLE and DRLE produce samples from their respective global minimisers, while DE produces a distribution which—-in accordance with Theorem 5.4.1—does not correspond to the global minimiser of $Q \mapsto \int \ell(\theta)\, dQ(\theta)$ over $\mathcal{P}(\mathbb{R}^J)$ (which is given as Dirac measure located at $\theta = -2$).

FD-GVI vs ID-GVI: Figure 5.3 illustrates two aspects. First, the effect of regularisation for DLE and

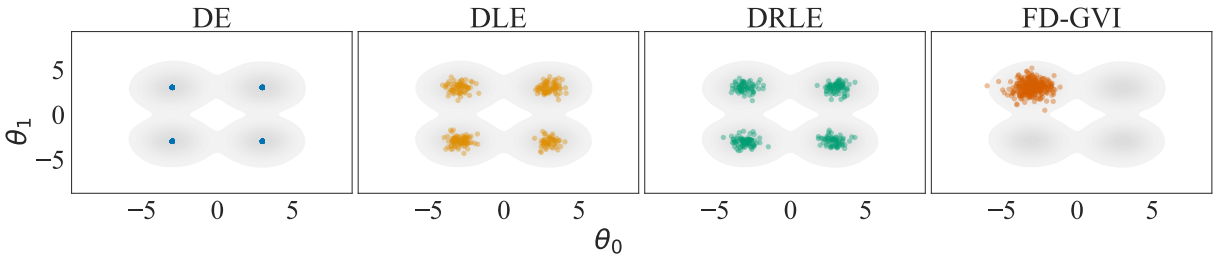


Figure 5.3: We generate $N_E = 300$ particles from DE, DLE, DRLE and FD-GVI with Gaussian parametrisation. The multimodal loss ℓ is plotted in grey and the particles of the different methods are layered on top. The prior in this example is flat, i.e. $\log p$ and μ_P are constant. The initialisation Q_0 is standard Gaussian. See Figure C.1 in the appendix for an alteration of this setting with only four particles.

	KIN8NM	CONCRETE	ENERGY	NAVAL	POWER	PROTEIN	WINE	YACHT
DE	$\mathbf{0.33_{\pm 0.1}}$	$6.10_{\pm 0.3}$	$2.83_{\pm 0.2}$	$-0.40_{\pm 0.3}$	$\mathbf{13.70_{\pm 2.6}}$	$\mathbf{11.22_{\pm 2.2}}$	$14.65_{\pm 1.9}$	$2.20_{\pm 0.4}$
DLE	$13.25_{\pm 4.3}$	$\mathbf{5.11_{\pm 0.2}}$	$\mathbf{2.43_{\pm 0.1}}$	$3.46_{\pm 2.4}$	$13.87_{\pm 2.3}$	$43.20_{\pm 12.5}$	$13.73_{\pm 1.4}$	$\mathbf{1.64_{\pm 0.1}}$
DRLE	$0.46_{\pm 0.1}$	$8.30_{\pm 0.6}$	$4.01_{\pm 0.3}$	$\mathbf{-3.04_{\pm 0.2}}$	$23.21_{\pm 2.0}$	$48.80_{\pm 2.1}$	$\mathbf{7.13_{\pm 0.6}}$	$7.80_{\pm 2.7}$

Table 5.1: Table compares the average (Gaussian) negative log likelihood in the test set for the three ID-GVI methods on some UCI-regression data sets [Lichman, 2013]. We observe that no method consistently outperforms any of the others.

DRLE is that particles spread out around the local minima. In comparison, DE particles fall directly into the local minima. Second, FD-GVI (with Gaussian parametric family) leads to qualitatively poorer approximations of Q^*. This is because the ID-GVI methods explore the whole space $\mathcal{P}_2(\mathbb{R}^J)$, whilst FD-GVI is limited to learning a unimodal Gaussian.

DEs vs D(R)LEs, and why finite N_E matters: Table 5.1 compares DE, DLE and DRLE on a number of real world data sets, and finds a rather random distribution of which method performs best. This seems to contradict our theory, and suggests there is essentially no difference between regularised and unregularised ID-GVI. What explains the discrepancy? Essentially, it is the fact that N_E is not only finite, but much smaller than the number of minima found in the loss landscape of deep learning. In this setting, each particle moves into the neighbourhood of a well-separated single local minimum and typically never escapes, even for very large T. We illustrate this in Figure 5.4 with a toy example. We choose a uniform prior P and initialisation Q_0 and the loss $\ell(\theta) := -|\sin(\theta)|$, $\theta \in [-1000\pi, 1000\pi]$, which has 2000 local minima. Correspondingly Q^* will have many local modes for all methods. Note that ∇V is the same for all approaches since $\log p$ and μ_P are constant. The difference between the methods boils down to repulsive and noise effects. However, these noise effects are not significant if each particle is stuck in a single mode: the particles will bounce around their local modes, but not explore other parts of the space. This implies that they will not improve the approximation quality of Q^*. Note that this problem is a direct parallel to multi-modality—a well-known problem for Markov Chain Monte Carlo methods [see e.g. Syed et al., 2022].

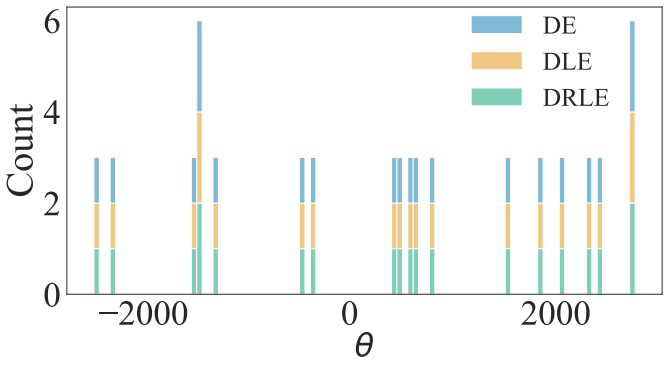


Figure 5.4: We generate $N_E = 20$ samples from the three infinite-dimensional gradient descent procedures discussed in Section 5.4. The x-axis shows the location of the particles after training. Since the same initialisation $\theta_n(0)$ for all methods is chosen we can observe that particles fall into the same local mode. Further, 16/20 particles are alone in their respective local modes and the location of the particles varies only very little between the different methods (which is why they are in the same bucket in the above histogram).

5.6 Conclusion

In this paper, we used infinite-dimensional gradient descent via Wasserstein gradient flows (WGFs) [see e.g. Ambrosio et al., 2005] and the lens of generalised variational inference (GVI) [Knoblauch et al., 2019] to unify a collection of existing algorithms under a common conceptual roof. Arguably, this reveals the WGF to be a powerful tool to analyse ensemble methods in deep learning and beyond. Our exposition offers a fresh perspective on these methodologies, and plants the seeds for new ensemble algorithms inspired by our theory. We illustrated this by deriving a new algorithm that includes a repulsion term, and use our theory to prove that ensembles produced by the algorithm converge to a global minimum. A number of experiments showed that the theory developed in the current paper is useful, and showed why the performance difference between simple deep ensembles and more intricate schemes may not be numerically discernible for loss landscapes with many local minima.

Statement of Authorship for joint/multi-authored papers for PGR thesis

Title of Paper	A Rigorous Link between Deep Ensembles and (Variational) Bayesian Methods
Publication Status	published
Publication Details	Veit D. Wild, Sahra Ghalebikesabi, Dino Sejdinovic and Jeremias Knoblauch. "A Rigorous Link between Deep Ensembles and (Variational) Bayesian Methods." Advances in Neural Information Processing Systems (NeurIPS), 2023.

Student Confirmation

Student Name:	Veit David Wild		
Contribution to the Paper	• Derivation of all theoretical results in the paper • Writing of the manuscript • Assistance in implementation of the algorithm		
Signature	V.Wild	Date	28.01.2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Dino Sejdinovic		
Supervisor comments			
Signature	Dino Sejdinovic	Date	30 January 2024

This completed form should be included in the thesis, at the end of the relevant chapter.

6 | Bayesian Inference in Function Space via the Wasserstein Gradient Flow

This chapter is unpublished and written in manuscript style.

Abstract

We propose a scalable inference algorithm for Bayes posteriors defined on a reproducing kernel Hilbert space (RKHS). Given a likelihood function and a Gaussian random element representing the prior, we obtain the corresponding Bayes posterior measure Π^* as the stationary distribution of an RKHS-valued Langevin diffusion. We approximate the infinite-dimensional Langevin diffusion via a projection onto the first M components onto the spectral basis of the covariance operator. Exploiting the thus obtained approximate posterior for these M components, we perform inference for Π^* by relying on the law of total probability and a sufficiency assumption. The resulting method scales as $O(M^3 + JM^2)$, where J is the number of samples produced from the posterior measure Π^*. Interestingly, the algorithm as a special case recovers the posterior arising from the sparse variational Gaussian process (SVGP) [Titsias, 2009a]–owed to the fact that the sufficiency assumption underlies both methods. However, whereas the SVGP posterior is parametrically constrained to be a Gaussian process, our method is based on a variational family that can freely explore the space of all probability measures on $\mathbb{R}^M$. Due to this additional flexibility, we can show that our method is close to the optimal M-dimensional variational approximation of the Bayesian posterior Π^* for convex and Lipschitz continuous negative log likelihoods extending the optimality of SVGP beyond the case of Gaussian likelihhods.

6.1 Introduction

Bayesian inference in function spaces crucially relies on the construction of a prior distributions on an infinite-dimensional function space. Gaussian processes (GPs) are the standard tool deployed for this purpose, and have proven a valuable framework for principled uncertainty quantification over the past few decades [Rasmussen and Williams, 2006]. A notable shortcoming of GPs is their cubic computational cost with respect to data size, and their intractability for the case of non-Gaussian likelihoods. Numerous approaches have been proposed to overcome these challenges [see e.g. Gneiting, 2002, Quiñonero-Candela and Rasmussen, 2005, Chalupka et al., 2013, Wilson and Nickisch, 2015, Gardner et al., 2018, Wang et al., 2019, Liu et al., 2020]. Amongst them, sparse variational Gaussian processes (SVGPs) [Titsias, 2009a] arguably are considered to be the gold standard for scalable GP approximations. SVGPs rely on the introduction of inducing features that are assumed to follow a Gaussian distribution with learnable variational parameters, and which can be interpreted as an approximate low-dimensional summary of the posterior GP. To learn these variational parameters, one then typically performs gradient based maximisation of the Evidence Lower Bound (ELBO) [see e.g. Hensman et al., 2013].

While SVGPs often perform surprisingly well in practice, their approximation quality is inadequate whenever the Gaussianity imposed upon the inducing points is too restrictive. In the hope of overcoming this shortcoming, our paper explores the possibility of performing Bayesian inference by using a version of gradient descent *directly* in the space of probability measures—rather than on variational parameters indexing an family of approximating measures [cf. Wild et al., 2023]. We achieve this via the Wasserstein Gradient Flow (WGF) [Otto, 2001, Ambrosio et al., 2005], which is a natural analogue for gradient descent over the space of probability measures. Implementing the WGF for our problem requires a reformulation of the Bayes posterior as an infinite-dimensional optimisation problem on the space of probability measures [Knoblauch et al., 2019, Wild et al., 2022].

To derive the WGF, we need a deep and precise understanding of the function space H on which our prior and therefore posterior is defined. This requirement forces us to incorporate the Function space H explicitly into our model of a random function which naturally leads to the structure of a Gaussian random element [Van Neerven, 2008, Wild and Wynne, 2022]. Simply put a Gaussian random element is a Gaussian process with known sample path properties [see e.g. Wild et al., 2022] and they therefore provide the perfect framework for calculations in infinite-dimensional analysis.

Having chosen a GRE on a suitable (Hilbert) function space H, we implement the WGF, which leads to a Langevin stochastic differential equation (SDE) evolving in H whose stationary distribution corresponds to the targeted Bayes posterior Π^*. Intriguingly, since we rely on Fréchet derivatives to represent this

SDE, we find that the underlying function space *has* to be an RKHS $H = H_k$ associated to some kernel function k. Since elements of H_k can generally not be represented numerically without approximation, we project this SDE in H into a finite-dimensional representation corresponding to the first M coefficients of the Kosambi–Karhunen–Loève expansion. The result is a standard M-dimensional Langevin diffusion with a well-understood limiting measure. Given this, the key insight is now to solve the inverse problem: by probabilistically mapping back from the limiting distribution of the projected SDE to the limiting distribution of infinite-dimensional SDE on H, we can draw approximate samples from the Bayes posterior Π^*. This gives rise to an algorithm we call Projected Langevin Sampling (PLS), which produces posterior inferences at computational cost of order $\mathcal{O}(M^3 + JM^2)$ for $J \in \mathbb{N}$ posterior samples, and which is applicable to generic likelihoods.

Perhaps surprisingly, our algorithm is both a sampling and a variational inference algorithm that generalises SVGPs in a particular way. Specifically, we can formally show that PLS samples from a variational approximation that is strictly more expressive than the corresponding variational posterior constructed with SVGP methods: while SVGPs force a the variational distribution over their M inducing features to be distributed according to a Gaussian measure, the variational family used within PLS is *not* parametrically constrained in any way, and simply given as the set of all probability measures over $\mathbb{R}^M$. As we show in our theoretical results, this allows PLS to target a strictly better posterior approximation than SVGPs.[1] We also find that for the special case of Gaussian likelihoods, both PLS and SVGPs target the same (Gaussian) approximate posterior measure. Unlike SVGPs, PLS can also be used for arbitrary likelihoods without requiring their integrals with respect to variationally parameterised Gaussian measures to be approximated via Markov chain Monte Carlo or be tractable.

6.2 Optimisation-centric Perspectives on Bayesian Inference

Bayesian inference in function spaces assigns a prior for an unknown function f linked to $N \in \mathbb{N}$ observations $y_{1:N} = (y_1, \ldots, y_N) \in \mathbb{R}^N$ via a likelihood $p(y_{1:N}|f)$[2]. The best known example is Gaussian process regression [Rasmussen and Williams, 2006] where the prior for f is a Gaussian process (GP), and the log-likelihood is given by

$$\log p(y_{1:N}|f) = \sum_{n=1}^{N} \log \mathcal{N}\big(y_n; f(x_n), \sigma^2\big).$$

[1] This holds under the natural comparison where the M coefficients of the Kosambi–Karhunen–Loève expansion are now used as the SVGP's inducing features.

[2] Here, $p(y_{1:N}|f)$ is called a likelihood function if $(y_{1:N}, f) \in \mathbb{R}^N \times H \to p(y_{1:N}|f) \in [0, \infty)$ is product-measurable, and if $\int_{\mathbb{R}^N} p(y_{1:N}|f)\, d\mu(y_{1:N}) = 1$ for all fixed $f \in H$, and for μ a base measure on $\mathbb{R}^N$—such as the counting measure or Lebesgue measure.

Here, $\sigma^2 > 0$ is observation noise and $x_{1:N} = (x_1, \ldots, x_N) \in \mathcal{X}^N$ are known features. Given the prior for f and this likelihood, the goal is to find the posterior for f given the observations $y_{1:N} \in \mathbb{R}^N$.

In this work our prior is modelled as a GRE F with mean zero and covariance operator $C : H \to H$, where H is the underlying Hilbert space of functions (see Appendix D.4 for an introduction to GREs). We write $F \sim \mathcal{N}(0, C)$ for a mean zero GRE with covariance operator C. Notice that the covariance operator C plays a similar role to the covariance function or kernel of a Gaussian process and allows us to incorporate prior knowledge about the unknown function into our model.

Denoting by Π the Gaussian measure associated with the GRE $F \sim \mathcal{N}(0, C)$, and by $\ell(f) := -\log p(y_{1:N}|f)$, $f \in H$, an arbitrary negative log likelihood function, it is well-known that the Bayes posterior, denoted Π^*, is the solution to a particular optimisation problem, both for the parametric and the non-parametric case [e.g., Theorem 1 in Knoblauch et al., 2019, Wild et al., 2022, Appendix A.1]. Indeed, the Bayes posterior Π^* can be written as

$$\Pi^* = \operatorname*{arg\,min}_{Q \in \mathcal{P}(H)} \underbrace{\int \ell(f)\, dQ(f) + \mathrm{KL}(Q, \Pi)}_{=L(Q)}, \tag{6.1}$$

where $\mathcal{P}(H)$ denotes the space of all probability measures on H, and $\mathrm{KL}(Q, \Pi)$ denotes the Kullback-Leibler divergence between the measures Q and Π.

This optimisation-centric perspective has many methodological and theoretical uses. For example, it shows that conventional variational inference relates to full Bayesian inference like constrained to unconstrained optimisation [see Theorem 2 in Knoblauch et al., 2019]. To see this, simply note that conventional variational inference consists in choosing a family of measures $\mathcal{Q} \subset \mathcal{P}(H)$ that has a finite-dimensional Euclidean parameterisation, and then computing the variational Bayes posterior as $\widehat{Q} = \operatorname{arg\,min}_{Q \in \mathcal{Q}} L(Q)$ through stochastic gradient descent [e.g. Titsias and Lázaro-Gredilla, 2014] or other optimisation techniques. In the specific context of Gaussian processes (GPs), such variational Bayesian approaches were pioneered by Titsias [2009a], clarified in Matthews et al. [2016], and further developed for Gaussian Random Elements (GREs) in Banach spaces by Wild and Wynne [2022].

In the current paper, we develop methodology that refrains from the conventional variational Bayes approach: instead of choosing a parameterised family of distributions $\mathcal{Q}$ and performing gradient descent for its finite-dimensional parameters, we instead perform gradient descent *directly* in the infinite-dimensional space $\mathcal{P}(H)$. Unlike for parameterised families $\mathcal{Q}$, this also for more flexibility and typically results in better posterior approximations [Wild et al., 2023]. In the context of SVGPs the variational family is constructed via inducing features, which are assumed to be Gaussian. Consequently, the resulting

posterior approximation is always a Gaussian process even if the true posterior deviates significantly from Gaussianity.

We therefore propose to circumvent the shortcomings of a restrictive variational family by implementing Gradient descent directly in the space of probability measures for the objective L in (6.1) via the Wasserstein gradient flow. This added flexibility is reflected in the fact that our method is provably close to optimal for many likelihoods of practical interest.

6.3 Wasserstein Gradient Flow (WGF) for Functional Inference

Before we can define the WGF on (6.1), we first define the prerequisite notations and objects. To this end, let $\mathcal{B}(H)$ be the Borel σ-algebra on H, and define

$$\mathcal{P}_2(H) := \left\{ \mu : \mathcal{B}(H) \to [0,1] \,\middle|\, \int_H ||u||^2 \, d\mu(u) < \infty \right\}$$

as the set of Borel probability measures on H with finite second moment. We can equip $\mathcal{P}_2(H)$ with the 2-Wasserstein metric distance, which for $\Pi, Q \in \mathcal{P}_2(H)$ is defined as

$$W_2(\Pi, Q)^2 := \inf \left\{ \int_{H \times H} ||h - h'||^2 \, d\gamma(h, h') : \gamma \in \mathcal{C}(\Pi, Q) \right\}.$$

Here, $\mathcal{C}(\Pi, Q) \subset \mathcal{P}_2(H \times H)$ is the set of all probability measure on $\mathcal{B}(H \times H)$ whose marginals are Π and Q, so that $\gamma(A \times H) = \Pi(A)$ and $\gamma(H \times B) = Q(B)$ for all $A, B \in \mathcal{B}(H)$ [cf. Ambrosio et al., 2005, Chapter 6].

6.3.1 Gradient Descent in $\mathcal{P}_2(H)$

Because the prior Π is a Gaussian measure, it is straightforward to show that the posterior is guaranteed to have a finite second moment, so that $\Pi^* \in \mathcal{P}_2(H)$ (cf. Appendix D.3 for a proof). Therefore, we can use the functional $Q \mapsto L(Q)$ in (6.1) to find the minimiser Π^* by performing a type of gradient descent in $\mathcal{P}_2(H)$. In particular and as outlined in Chapter 2 of Ambrosio et al. [2005], we can pick a starting point $Q_0 \in \mathcal{P}_2(H)$, and then iteratively update according to

$$Q_{k+1} := \operatorname*{arg\,min}_{Q \in \mathcal{P}_2(H)} \left\{ L(Q) + \frac{1}{2\eta} W_2(Q, Q_k)^2 \right\},$$

for $k \in \mathbb{N}$ and $\eta > 0$ a sufficiently small step-size. If L is sufficiently regular and we let $\eta \to 0$, the continuously indexed family $\left(Q_{\lfloor t/\eta \rfloor}\right)_{t \geq 0}$ converges to a limit $\left(Q(t)\right)_{t \geq 0} \subset \mathcal{P}_2(H)$ with $Q(0) = Q_0$ [Ambrosio et al., 2005, Chapter 11.1.3]. This evolution of measures $\left(Q(t)\right)_{t \geq 0}$ is called the Wasserstein

gradient flow (WGF) for L starting at Q_0. It has obvious parallels with gradient descent, and is directly interpretable as its infinite-dimensional analogy in $\mathcal{P}_2(H)$: at every infinitesimally small step forward in time t, $Q(t)$ moves in the direction amounting to the biggest decrease in the value for $Q \mapsto L(Q)$.

6.3.2 Following the WGF via the Langevin SDE in Hilbert Space

Following Theorem 8.3.1 in Ambrosio et al. [2005], we know that the WGF $\big(Q(t)\big)_{t>0}$ for functional $Q \mapsto L(Q)$ in (6.1) satisfies for all test functions [3] $\varphi : [0,T] \times H \to \mathbb{R}$ the equation

$$\int_0^T \int_H \partial_t \varphi(t,f) - \big\langle \nabla_W L[Q(t)](f), D\varphi(t,f) \big\rangle \, dQ_t(f)\, dt = 0, \tag{6.2}$$

where $D\varphi(t,f)$ denotes the Fréchet derivative of φ with respect to f, and $\nabla_W L[Q] : H \to H$ is the so-called *Wasserstein gradient* of L evaluated at $Q \in \mathcal{P}_2(H)$. In Theorem D.2.3 (cf. Appendix D.2), we show that this expression is given by

$$\nabla_W L[Q](f) = D\ell(f) + D(\log q)(f) = D\ell(f) + \frac{Dq(f)}{q(f)}, \tag{6.3}$$

where $q := dQ/d\Pi : H \to \mathbb{R}$ is the Radon-Nikodym derivative of Q with respect to the prior measure Π. Plugging (6.3) into (6.2), we can identify a stochastic process $\big(F(t)\big)_{t>0}$ with $F(t) \in H$ for which $F(t) \sim Q(t)$ for all $t \geq 0$ (cf. Theorem D.2.2 in Appendix D.2). In particular, we find that the solution to the infinite-dimensional version of the Langevin Stochastic Differential Equation (SDE)

$$dF(t) = -\big(D\ell\big(F(t)\big) + C^{-1}F(t)\big)\, dt + \sqrt{2} dW(t), \tag{6.4}$$

where $\big(W(t)\big)_{t\geq 0}$ is a cylindrical Brownian motion in H and $C^{-1} : \mathrm{Im}(C) \subset H \to H$ is the inverse of C[4], follows the Wasserstein gradient flow. This SDE is introduced in Hairer et al. [2005], and the existence, uniqueness and ergodic properties of its solution are discussed in Hairer et al. [2007b]. Hairer et al. [2011] provides a more accessible treatment of the subject, and Da Prato and Zabczyk [2014] provides a more general discussion of SDEs in Hilbert spaces.

For our purposes, we can thankfully gloss over many of the technical aspects of dealing with this infinite-dimensional SDE. Instead, our main reason for introducing (6.4) will be its use as the driving engine for our methodology. To this end, it is instructive to note the parallels with the finite-dimensional case: if

[3] A sufficiently large class of test functions are the smooth cylindrical function, denoted $\mathrm{Cyl}\big((H \times [0,T]\big)$ [Ambrosio et al., 2005, Definition 5.1.11]. A function φ is contained in $\mathrm{Cyl}\big((H \times [0,T]\big)$ if and only if there exists $N \in \mathbb{N}$ and orthonormal vectors $v_1, \ldots, v_N \in H$ such that $\varphi(f) = \psi\big(\langle f, v_1\rangle, \ldots, \langle f, v_N\rangle, t\big)$ where $\psi : \mathbb{R}^N \times [0,T] \to \mathbb{R}$ is a smooth function with compact support.

[4] The operator C^{-1} is unbounded and given by $C^{-1}f = \sum_{n=1}^{\infty} \lambda_n^{-1} \langle f, e_n \rangle e_n$ where $\{\lambda_n, e_n\}_{n=1}^{\infty}$ are the eigenvalue-eigenvector pairs associated with the self-adjoint trace class covariance operator C.

we consider (6.1) for $H = \mathbb{R}^J$, we recover a Bayes posterior defined on a finite-dimensional parameter $\theta \in \mathbb{R}^J$. Writing the associated likelihood function as $p(y_{1:N}|\theta)$, and the corresponding Gaussian prior density as $p(\theta) = \mathcal{N}(\theta; 0, \Sigma)$, the resulting WGF—now defined on $\mathcal{P}(\mathbb{R}^J)$—yields a measure evolution that corresponds to the well-known finite-dimensional Langevin SDE

$$d\theta(t) = -\left(\nabla \ell_N\big(\theta(t)\big) + \Sigma^{-1}\theta(t)\right) dt + \sqrt{2} d\beta(t), \tag{6.5}$$

where $\ell_N(\theta) := -\log p(y_{1:N}|\theta)$ and $\big(\beta(t)\big)_{t\geq 0}$ is the standard Brownian motion in $\mathbb{R}^J$ [Jordan et al., 1998, Otto, 2001]. This finite-dimensional Langevin SDE is the basis of many classical algorithms for Bayesian computation since its stationary distribution for $\theta(t)$ as $t \to \infty$ recovers the Bayes posterior [see e.g. Roberts and Tweedie, 1996, Welling and Teh, 2011]. In other words, and exactly as for the infinite-dimensional case, one way of motivating *why* this diffusion recovers the Bayes posterior is by interpreting it as the natural analogy of gradient descent for the (6.1) for the special case of $H = \mathbb{R}^J$. Just as (6.5) is the inspiration for many seminal algorithms for computing Bayes posteriors over finite-dimensional parameters, (6.4) can be used to generate samples from the posterior measure Π^* that solves the optimisation problem (6.1) with infinite-dimensional function spaces H. However, the infinite dimension of H introduces an additional complication: the Langevin SDE in (6.4) generally cannot easily be represented on a computer. Fortunately, we can overcome this with parsimonious approximations (cf. Section 6.5).

6.4 Choosing Parameter Space H and Prior Π

So far, our developments are valid for likelihoods and GREs on general Hilbert spaces. In particular, the Langevin SDE in (6.4) seemingly allows us to perform posterior inference for any parameter $f \in H$ so long as it lives in a Hilbert space. As we will explain next however, the seemingly innocuous requirement that ℓ be Fréchet differentiable will necessitate the assumption that H is a reproducing kernel Hilbert space (RKHS) for virtually all likelihood functions of practical interest (cf. Appendix D.1.1 for basic properties of an RKHS).

6.4.1 The Inevitability of the RKHS

Implementing the Langevin SDE in (6.4) requires calculation of the Wasserstein gradient in (6.3), and therefore of the Fréchet derivative $D\ell(f)$ of $\ell : H \to \mathbb{R}$. For simplicity, consider losses which for a cost

function $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ can be written as

$$\ell(f) = \sum_{n=1}^{N} c\big(y_n, f(x_n)\big) + \text{ constant}, \tag{6.6}$$

where $c\big(y_n, f(x_n)\big)$ measures the discrepancy between an observation y_n and the prediction $f(x_n)$ for y_n, and where the constant does not depend on f. For example, the Gaussian likelihood is based on $c(y_n, f(x_n)) = \frac{1}{2\sigma^2}\big(y_n - f(x_n)\big)^2$. Another example is the Bernoulli likelihood for binary classification, which is

$$p(y_n|f) = \phi\big(f(x_n)\big)^{y_n} \cdot \Big(1 - \phi\big(f(x_n)\big)\Big)^{1-y_n}. \tag{6.7}$$

In this case, $y_n \in \{0, 1\}$, so that the associated negative log likelihood $\ell(f)$ is of the form (6.6) with $c(y_n, f(x_n)) = -y_n \log \phi(f(x_n)) - (1 - y_n) \log(1 - \phi(f(x_n)))$. Here, $\phi : \mathbb{R} \to [0, 1]$ is a mapping that transforms latent functional outputs into probabilities such as the the logistic function $\phi_{\text{logistic}}(f(x_n)) = \big(1 + \exp(-f(x_n))\big)^{-1}$.

For any negative log likelihood with the form of (6.6), we can apply the chain rule and obtain the corresponding Fréchet derivative as

$$D\ell(f) = \sum_{n=1}^{N} (\partial_2 c)\big(y_n, f(x_n)\big) D(s_n)(f), \tag{6.8}$$

where we have written $s_n(f) := f(x_n)$ as the point-wise evaluation functional for $f \in H$ at x_n, $\partial_2 c$ as the derivative of c with respect to its second component, and $D(s_n)$ as the Fréchet derivative of s_n. Inspecting the expression, it is now clear that we need to assume that all s_n are Fréchet differentiable and therefore continuous. However, demanding continuity of all pointwise evaluation functionals $s_n(f)$ for all $f \in H$ is in fact equivalent to H being an RKHS [Berlinet and Thomas-Agnan, 2004, Theorem 1].

In summary, for likelihood functions based on pointwise evaluations of f as in (6.6), the Wasserstein gradient in (6.3) needed to evolve the infinite-dimensional Langevin SDE in (6.4) can only be implemented if H is an RKHS. Throughout the remainder of the paper, we therefore take H to be an RKHS H_k associated with the reproducing kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Thanks to the widely celebrated reproducing property of RKHS functions, we now have that $s_n(f) = \langle f, k_n \rangle$ for $k_n = k(x_n, \cdot) \in H_k$, so that the required Fréchet derivative in (6.8) is $D(s_n) = k_n$, and $D\ell(f) = \sum_{n=1}^{N} (\partial_2 c)\big(y_n, f(x_n)\big) k_n$.

6.4.2 Covariance Operators and Gaussian Processes

The GRE prior $F \sim \mathcal{N}(0, C)$ is specified by a covariance operator $C : H_k \to H_k$ that encodes prior knowledge about the unknown function. In this paper, we specify C through a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a probability measure $\nu \in \mathcal{P}(\mathcal{X})$ as

$$Cf := \int k(\cdot, x')\, f(x')\, d\nu(x') \tag{6.9}$$

for all $f \in H_k$. In Appendix D.4, we show that constructing C in this way indeed leads to a valid covariance operator on H_k under mild assumptions on ν and k.[5] Every GRE in H_k naturally induces a Gaussian process (GP) through the definition $\widetilde{F}(x) := \langle F, k(x, \cdot)\rangle$ for all $x \in \mathcal{X}$. The resulting stochastic process $\widetilde{F}$ is a GP with kernel r (cf. Lemma D.4.2 in Appendix D.4) given by

$$r(x, x') := \int k(x, \xi) k(\xi, x')\, d\nu(\xi) \tag{6.10}$$

for all $x, x' \in \mathcal{X}$. This connection is practically useful, since it is well-understood how to specify interpretable GP function priors [see e.g. Rasmussen and Williams, 2006, Duvenaud, 2014]. Therefore, one can specify interpretable GRE priors via the GP induced by r.[6] The correspondence between C and r also elucidates the role of ν: The similarity between two points x and x' is measured by the product of their pairwise similarities with a third point ξ as $k(x, \xi)k(\xi, x')$, and this product is averaged by ν over ξ. This construction often leads to a poor prior when ν deviates too much from the data generating process $\mathrm{Law}[X_1]$ of $x_1, \ldots, x_N$ (see Appendix D.5 for a more extensive discussion). To this end, we choose $\nu = \mathrm{Law}[X_1]$, and estimate the kernel r via samples whenever necessary.

6.5 Posterior Inference via Projection

Having laid out how the GRE prior should be designed, we are ready to deploy the infinite-dimensional Langevin SDE in (6.4) to produce samples from Bayes posteriors over H_k that are specified via (6.1) and rely on GRE priors. Inspecting the SDE, it quickly becomes clear that there is a poignant practical issue with using it to produce samples from the posterior: since it evolves over an RKHS H_k, there is no numerically exact way of representing $\big(F(t)\big)_{t \geq 0}$. A natural approach to solving this problem relates to

[5] A GRE $F \sim \mathcal{N}(0, C)$ exists if and only if C is a positive, self adjoint, trace-class operator [Da Prato and Zabczyk, 2014].

[6] Note that the GP prior $F \sim \mathrm{GP}(0, k)$ would *not* define a prior on the chosen function space H_k: by Dricoll's Theorem [Driscoll, 1973], the sample paths of such a GP would be almost surely not be contained in H_k.

the representation of $F \sim \mathcal{N}(0, C)$ in the spectral basis as

$$F(x) = \sum_{m=1}^{\infty} \langle F, e_m \rangle e_m(x),$$

where $F^m := \langle F, e_m \rangle \sim \mathcal{N}(0, \lambda_m)$ independently with $\{\lambda_m, e_m\}_{m=1}^{\infty} \subset [0, \infty) \times H_k$ the eigenvalue-eigenfunction pairs with $\lambda_1 \geq \lambda_2 \geq \dots$ obtained from the spectral decomposition of the covariance operator C in (6.9). This motivates the approximation of F via its orthogonal projection onto the first M eigenfunctions given by

$$\Pr[F] := \sum_{m=1}^{M} F^m e_m. \tag{6.11}$$

Since C is a trace-class operator, this approximation is well-motivated: the spectral decay of λ_m asymptotically faster than $1/m$. Indeed, this lower bound is often too conservative, and λ_m decays much faster rates if the functions in H_k are sufficiently smooth. As Table 6.1 illustrates, the decay is even often exponential, which means that roughly speaking, the projection in (6.11) converges to $F(t)$ exponentially fast as $M \to \infty$. From a practical standpoint, the projection is interesting because it is finite-dimensional—so that we can now study the appropriately projected finite-dimensional SDE for $(F^m(t))_{t \geq 0}$ and $m = 1, 2, \dots M$ instead of the intractable infinite-dimensional SDE in (6.4).

Table 6.1: Spectral decay for different kernels k and input distributions $\nu \in \mathcal{P}(\mathbb{R}^D)$; taken from Burt et al. [2019] and containing results from Ritter et al. [1995].

kernel	input distribution ν	decay of λ_m
Squared Exponential	compact support	$\mathcal{O}\big(\exp(-\alpha \frac{m}{D} \log \frac{m}{D})\big)$
Squared Exponential	Gaussian	$\mathcal{O}\big(\exp(-\alpha \frac{m}{D})\big)$
Matérn $l + 1/2$	Uniform	$\mathcal{O}\big(m^{-2l-2} \log(m)^{2(D-1)(l+1)}\big)$

6.5.1 Finite-Dimensional Projection of the WGF in Hilbert Space

Using Itó's Rule, we derive the time evolution for $F^m(t) := \langle F(t), e_m \rangle \in \mathbb{R}$ as

$$dF^m(t) = -\Big(\sum_{n=1}^{N} (\partial_2 c)\big(y_n, F(t)(x_n)\big) e_m(x) + \frac{F^m(t)}{\lambda_m} \Big) dt + \sqrt{2} dB^m(t), \tag{6.12}$$

where $B^m(t) := \langle W(t), e_m \rangle$ are stochastically independent Brownian motions in $\mathbb{R}$. Since the exact values of $\{\lambda_m, e_m\}_{m=1}^{M}$ and $\{F(t)(x_n)\}_{n=1}^{N}$ are unknown in practice, we require one additional layer of approximation. To this end, we use Nyström's method to estimate the M largest eigenvalue-eigenfunction pairs $\{\widehat{\lambda}_m, \widehat{e}_m\}_{m=1}^{M}$ (cf. Appendix D.7). Given $\{\widehat{\lambda}_m, \widehat{e}_m\}_{m=1}^{M}$, a natural estimate for $F(t)(x_n)$ then

follows from the estimated projection

$$\widehat{\text{Proj}}[F(t)] := \sum_{m=1}^{M} \underbrace{\langle F(t), \widehat{e}_m \rangle}_{=:\widehat{F}^m(t)} \widehat{e}_m \tag{6.13}$$

via $\widehat{\text{Proj}}[F(t)](x_n) = \sum_{m=1}^{M} \widehat{F}^m(t)\widehat{e}_m(x_n)$. Substituting these approximations into (6.12) leads to the SDE $\widehat{F}^{1:M}(t)$ whose components evolve as

$$d\widehat{F}^m(t) = -\left(\sum_{n=1}^{N} (\partial_2 c)\Big(y_n, \widehat{\text{Proj}}[F_t](x_n)\Big)\widehat{e}_m(x) + \frac{\widehat{F}^m(t)}{\widehat{\lambda}_m}\right) dt + \sqrt{2} dB^m(t). \tag{6.14}$$

This finally provides us with a fully tractable diffusion that can be simulated via Euler-Maruyama discretisation.

A naive methodology would now be based on the following observations: first, it is clear that the limiting distribution of $\big(F(t)\big)_{t\geq 0}$ is the Bayes posterior Π^*. Next, it is reasonable to expect that as $t \to \infty$, the distribution of $\widehat{F}^{1:M}(t) = (\widehat{F}^1(t), \dots \widehat{F}^M(t))^\top$ evolved according to (6.14) is such that $\widehat{\text{Proj}}[F(t)]$ constructed from these components follows a distribution that approximates Π^* as $t \to \infty$. More formally, defining $\widehat{\tau}_\infty \in \mathcal{P}(\mathbb{R}^M)$ as the limiting distribution of the SDE $(\widehat{F}^{1:M}(t))_{t\geq 0}$, we could pursue an approximate sampling algorithm for Π^* based on drawing samples from $\widehat{\tau}_\infty$, and then constructing a projection as in (6.13) based on these samples.

However, we can construct a much smarter approximation with $\widehat{\tau}_\infty$. In particular, it is reasonable to assume that $\widehat{\tau}_\infty \approx \widehat{\Pi}^{*,1:M}$, where $\widehat{\Pi}^{*,1:M}$ is the Bayes posterior distribution of $\widehat{F}^{1:M}|y_{1:N}$. Formally, we may obtain it as $\widehat{\Pi}^{*,1:M} = \phi\#\Pi^*$, where $\#$ denotes the pushforward operator and $\phi(f) := \langle f, \widehat{e}^{1:M}\rangle = (\langle f, \widehat{e}_1\rangle, \dots, \langle f, \widehat{e}_M\rangle)^\top$ is the function $\phi : H_k \to \mathbb{R}^M$ that maps $f \in H_k$ into the M largest estimated components used for the projection in (6.11). Since $\widehat{F}^{1:M}|y_{1:N} \sim \widehat{\Pi}^{*,1:M}$, a simple question can lead us to a much better approximation: given $\widehat{\Pi}^{*,1:M} = \phi\#\Pi^*$ is obtained by summarising $F|y_{1:N} \sim \Pi^*$ into an M-dimensional random quantity $\phi(F) = \widehat{F}^{1:M}$, how can we solve the inverse problem?

More specifically, what can we infer about the functional posterior $F|y_{1:N}$ if we know the posterior $\widehat{F}^{1:M}|y_{1:N}$ of the M coefficients?

As we show next, we can answer this question using the law of total probability and a sufficiency assumption commonly used in the so-called sparse variational GP (SVGP).

6.5.2 Approximate Projections for Exact Bayesian Inference

We want to use that $\widehat{\tau}_\infty \approx \widehat{\Pi}^{*,1:M} = \text{Law}[\widehat{F}^{1:M}|y_{1:N}]$ to approximate the predictive posterior distribution $F(x^*_{1:N_*})|y_{1:N}$ for arbitrary inputs $x^*_{1:N_*} \in \mathcal{X}^{N_*}$. To this end, we define the point-wise evaluation functional $\text{eval}[x^*_{1:N_*}](f) := f(x^*_{1:N_*})$ for $f \in H_k$, and the pushforward operator $\#$. With this, we write $\mathbb{P}(F(x^*_{1:N_*}) \in A \mid y_{1:N}) = \big(\text{eval}[x^*_{1:N_*}]\#\Pi^*\big)(A)$ for all measurable subsets $A \subset \mathbb{R}^{N_*}$. Noting further that $\widehat{\Pi}^{*,1:M}(A') = \mathbb{P}(\widehat{F}^{1:M} \in A'|y_{1:N})$ for all measurable subsets $A' \subset \mathbb{R}^M$, we find $\Pi^*(B) = \int \mathbb{P}(F \in B \mid y_{1:N}, \widehat{F}^{1:M} = u) d\widehat{\Pi}^{*,1:M}(u)$ for $B \in \mathcal{B}(H_k)$ by the law of total probability, and finally obtain the elementary identity

$$\big(\text{eval}[x^*_{1:N_*}]\#\Pi^*\big)(A) = \int_{\mathbb{R}^M} \mathbb{P}(F(x^*_{1:N_*}) \in A \,|\, y_{1:N}, \widehat{F}^{1:M} = u)\, d\widehat{\Pi}^{*,1:M}(u). \tag{6.15}$$

Given this, if we could sample from $\mathbb{P}(F(x^*_{1:N_*}) \in \cdot \,|\, y_{1:N}, \widehat{F}^{1:M} = u)$ in (6.15), then we could also generate samples that are approximately distributed according to the posterior predictive $F(x^*_{1:N_*})|y_{1:N} \sim \text{eval}[x^*_{1:N_*}]\#\Pi^*$. This holds since $\widehat{\tau}_\infty \approx \widehat{\Pi}^{*,1:M}$, so that it is straightforward to produce approximate samples from $\widehat{\Pi}^{*,1:M}$ via forward-simulation of (6.14), and to propagate them through the conditional probability in (6.15).

While we generally do not know the conditional exactly, we do know it for the special case where $p(y_{1:N}|f)$ is a Gaussian likelihood: now, the joint distribution of $\big(F(x^*_{1:N_*}), Y_{1:N}, \widehat{F}^{1:M}\big)$ is Gaussian with a known covariance matrix (cf. Appendix D.1). However, even for this extremely limited setting, the resulting sampling algorithm would be computationally infeasible. In particular, it would require inversion of an $(N+M) \times (N+M)$ matrix, meaning that drawing J samples would now scale as $\mathcal{O}((N+M)^3 + J(N+M)^2)$. Notably, this even exceeds the $\mathcal{O}(N^3)$ computational burden of exact Bayesian inference using GPs.

Our approach reduces computational cost via an approximation inspired by the way in which sparse variational Gaussian processes (SVGPs) use $M \ll N$ so-called inducing points [see Titsias, 2009a]. By assuming that conditioning on M inducing points rather than the full data set, SVGPs reduce the computational effort required for GP inference to $\mathcal{O}(M^3)$. For our method, we use a simplifying assumption of similar character: in particular, we assume that knowledge of the first M estimated components $\widehat{F}^{1:M}$ of the orthogonal projection in (6.13) is sufficient for inference, so that knowing $y_{1:N}$ would provide a negligible amount of additional information. We call this the *sufficiency condition*, and formally express it as

$$\mathbb{P}(F(x^*_{1:N_*}) \in A \,|\, y_{1:N}, \widehat{F}^{1:M} = u) \approx \mathbb{P}(F(x^*_{1:N_*}) \in A \,|\, \widehat{F}^{1:M} = u) \;\text{ for all } u \in \mathbb{R}^M. \tag{6.16}$$

Importantly, the right hand side of this relation is conditionally Gaussian, scales as $\mathcal{O}(M^3 + JM^2)$, and can be sampled from by, for example, Matheron's Rule [Journel and Huijbregts, 1976, Wilson et al., 2020]. In direct parallel to the inducing point framework of SVGPs, while the approximate relationship in (6.16) is only exact when $N = M$, experimental findings demonstrate that it remains an extremely good approximation even if $M \ll N$. The interpretation of (6.16) is straightforward: in terms of the projection in (6.13), it reflects the belief that knowledge of the first M terms in the spectral representation of $F \sim \Pi$ are sufficiently informative for inference, and that ignoring the remaining terms in the expansion leads to a negligible error. It is this analogy to the use of sufficient statistics in classical statistical methodology that gives the condition in (6.16) its name. Plugging (6.16) into (6.15) now gives rise to our final approximation $\widehat{\Pi} \in \mathcal{P}_2(H_k)$ of Π^*, which for any $B \in \mathcal{B}(H_k)$ is defined as $\widehat{\Pi}(B) := \int \mathbb{P}(F \in B \,|\, \widehat{F}^{1:M} = u) d\widehat{\tau}_\infty(u)$. Based on $\widehat{\Pi}$, we can now approximate the posterior predictive in (6.15) via

$$\big(\,\mathrm{eval}[x^*_{1:N_*}]\#\widehat{\Pi}\big)(A) = \int \mathbb{P}(F(x^*_{1:N_*}) \in A \,|\, \widehat{F}^{1:M} = u)\, d\widehat{\tau}_\infty(u). \tag{6.17}$$

In Section 6.5.3, we describe in detail how to sample from this measure.

To summarise: We move from Π^* to $\widehat{\Pi}$ via two approximation $\pi_\infty \approx \widehat{\Pi}^{*,1:M}$, and the sufficiency condition in (6.16). Generally, these two approximations do not hold with equality. The one notable exception to this is the setting where both $M = N$, and $\nu = \frac{1}{N}\sum_{n=1}^N \delta_{x_n}$. For this special case, $\widehat{\tau}_\infty = \widehat{\Pi}^{*,1:N}$ and therefore $\widehat{\Pi} = \Pi^*$, so that the method proposed in Section 6.5.3 is an exact sampling algorithm for Π^*.[7]

6.5.3 Projected Langevin Sampling (PLS)

Whether we choose to conduct exact inference with $M = N$ or approximate inference with $M \ll N$, the sampling algorithm is the same. Drawing samples that are distributed according to $\mathrm{eval}[x^*_{1:N_*}]\#\widehat{\Pi}$ is straightforward and proceeds in two steps: first, we evolve J independent SDEs as in (6.14) to obtain exact draws $\widehat{F}_1^{1:M}, \widehat{F}_2^{1:M}, \ldots \widehat{F}_J^{1:M}$ from the distribution $\widehat{\tau}_\infty$. Second, we then use Matheron's Rule (cf. Appendix D.9) to convert them into samples from $\mathrm{eval}[x^*_{1:N_*}]\#\widehat{\Pi}$ (cf. Appendix D.9). In other words, we obtain samples by

1. Sampling the initial conditions of our J prequisite SDEs as $\widehat{F}_j^{1:M}(0) \sim \tau_0$ where $\tau_0 \in \mathcal{P}(\mathbb{R}^M)$ is a user-chosen initial distribution and $j = 1, 2, \ldots J$;

2. Using the Euler-Maruyama discretisation to simulate SDEs with initial condition $\widehat{F}_j^{1:M}(0)$ forward

[7] The fact that $\widehat{\tau}_\infty \neq \widehat{\Pi}^{*,1:N}$ can be understood by decomposing $\widehat{\tau}_\infty$ and $\widehat{\Pi}^{*,1:N}$ into their likelihood and prior components. When $M < N$, the likelihoods differ. For $M = N$, the likelihoods match, and the only difference comes from the prior, and specifically the estimated eigenvalues. If $M = N$ and we choose $\nu = \frac{1}{N}\sum_{n=1}^N \delta_{x_n}$ for the covariance operator C, then this difference also vanishes (cf. Appendix D.8).

until time T according to (6.14), thereby obtaining J approximate[8] samples $\widehat{F}_j^{1:M}(T) \sim \widehat{\tau}_\infty$, for $j = 1, 2, \dots J$;

3. Applying Matheron's rule [Journel and Huijbregts, 1976, Wilson et al., 2020] to implement the integration on the right of (6.1) for any arbitrary set of inputs $x^*_{1:N_*} \in \mathcal{X}^{N_*}$. Let $G_1, \dots, G_J \sim \mathcal{N}(0, C)$ be stochastically independent GRE and denote $G_j(x^*_{1:N_*}) = (G_j(x^*_1), \dots G_j(x^*_{N_*}))^\top$, $\langle G_j, \widehat{e}^{1:M} \rangle = (\langle G_j, \widehat{e}_1 \rangle, \dots \langle G_j, \widehat{e}_M \rangle)^\top$. By standard rules for GREs, we can sample $\big(G_j(x^*_{1:N_*}), \langle G_j, \widehat{e} \rangle\big)^\top \sim \mathcal{N}(0, R_{N_*,M})$ for a covariance matrix $R_{N_*,M}$ defined as,

$$R_{N_*,M} := \begin{bmatrix} r(x_{1:N_*}, x_{1:N_*}) & \widehat{e}^{1:M}(x_{1:N_*})^T \widehat{\Lambda}_M \\ \widehat{\Lambda}_M \widehat{e}^{1:M}(x_{1:N_*}) & \widehat{\Lambda}_M \end{bmatrix} \in \mathbb{R}^{(N_*+M)\times(N_*+M)} \tag{6.18}$$

which allows us to obtain the J approximate posterior samples $F_j(x^*_{1:N_*}) \sim \mathrm{eval}[x^*_{1:N_*}]\#\widehat{\Pi}$ for $j = 1, 2, \dots J$ as

$$F_j(x^*_{1:N_*}) = G_j(x^*_{1:N_*}) + \widehat{e}^{1:M}(x^*_{1:N_*})^\top \Big(\widehat{F}_j^{1:M}(T) - \langle G_j, \widehat{e}^{1:M} \rangle\Big).$$

Here, $\widehat{e}^{1:M}(x^*_{1:N_*}) \in \mathbb{R}^{M\times N_*}$ is the matrix whose entry at (m, n) is $\widehat{e}_m(x^*_n)$ and $\widehat{\Lambda}_M := \mathrm{diag}(\widehat{\lambda}_1, \dots, \widehat{\lambda}_M) \in \mathbb{R}^{M\times M}$ is the diagonal matrix with entries $\widehat{\lambda}_1, \dots, \widehat{\lambda}_M$.

Note that since each SDE in the first two steps above can be evolved without interaction, the entire sampling algorithm is embarrassingly parallel over the number of samples J. Thus, while a naive implementation would scale as $\mathcal{O}(M^3 + JM^2)$, parallelisation speeds things up significantly. For example, we found that the differences in computation time between $J = 1$ and $J = 100$ in our parallelised implementation were negligible.

A more detailed version of the algorithm can be found in Appendix D.14, which also includes further implementation details. Since the underlying inferential engine is the projection of the infinite-dimensional Langevin SDE in (6.4) into a finite-dimensional presentation as in (6.14), we will refer to this algorithm as *projected Langevin sampling* (PLS) throughout the remainder of the paper.

6.6 Theoretical Analysis of Projected Langevin Sampling (PLS)

So far, the motivation for our algorithm was guided from a methodological point of view: we started from an intractable inference problem, and found a way of producing an approximate algorithm for it. In line with this, we developed intuitions through heuristics and several layers of approximations. In

[8] This approximation is due to the finite amount of time T and the discretisation error of the SDE, and vanishes the finer the discretisation gets and the larger T becomes.

the remainder, we will demonstrate that the choices we made along the way are sound, and that our methodological developments can be rigorously justified. To do so, we first study the theoretically optimal approximation for Π^* over the class of approximations allowed to depend on the first M components of the orthogonal projection in (6.11) (cf. Theorem 6.6.1). Perhaps surprisingly, we obtain an exact expression for this approximation. Based on this, we study the quality difference between the theoretically optimal approximation and our proposal $\widehat{\Pi}$ in Theorem 6.6.2. The result is an explicit bound on this difference in terms of the eigenvalues $\{\lambda_m\}_{m>M}$ corresponding to the terms that were left out in the projection of (6.11), illustrating that the error depends on spectral decay of the eigenvalues $\{\lambda_m\}_{m=1}^{\infty}$ of the covariance operator C. In our last result, we use this bound to show that for the special case of Gaussian likelihoods, our approximation $\widehat{\Pi}$ coincides with both this optimal approximation and with SVGP posteriors (Lemma 6.6.3).

6.6.1 Assumptions and Notations

For the remainder, we will assume that we can rewrite $-\log p(y_{1:N}|f) =: \ell_N(f(x_{1:N}))$ for an appropriately defined function $\ell_N : \mathbb{R}^N \to \mathbb{R}$ that is allowed to depend on $y_{1:N}$. This is strictly more general than the form assumed for the derivation of the WGF in (6.6)[9], and more notationally convenient.

Throughout our theoretical developments, we further assume that we have access to the M-largest eigenvalue-eigenfunction pairs $\{\lambda_m, e_m\}_{m=1}^M$ of the covariance operator C. This will allow us to ignore the estimation error of $\{\widehat{\lambda}_m, \widehat{e}_m\}_{m=1}^M$ and simplify the already challenging mathematical arguments. While there generally will be estimation errors when Nyström's method is used, it is generally believed that they are small for the larger eigenvalue-eigenfunction pairs [cf. Section 4.3.2 in Rasmussen and Williams, 2006]. In some cases, the error is even exactly zero, since the decomposition $\{\lambda_m, e_m\}_{m=1}^M$ is actually also known exactly for a number of kernels k and input distributions ν [Zhu et al., 1997]). Based on this assumption, we now define τ_∞ as the limiting measure of the SDE (6.14) with $\{\widehat{\lambda}_m, \widehat{e}_m\}_{m=1}^M$ substituted for by $\{\lambda_m, e_m\}_{m=1}^M$.[10] Recalling that we defined $F^m = \langle F, e_m \rangle$ for $m = 1, \ldots, M$ and $F^{1:M} = (F^1, \ldots F^M)^\top$, we will now overload notation for $\widehat{\Pi}$ by writing the approximate posterior depending on τ_∞ (rather than $\widehat{\tau}_\infty$) as

$$\widehat{\Pi}(B) = \int \mathbb{P}(F \in B \,|\, F^{1:M} = u)\, d\tau_\infty(u), \tag{6.19}$$

for all $B \in \mathcal{B}(H_k)$. Throughout the remainder of Section 6.6, this is the definition of $\widehat{\Pi}$ our theoretical results are derived with respect to. While the differences between $\{\widehat{\lambda}_m, \widehat{e}_m\}_{m=1}^M$ and $\{\lambda_m, e_m\}_{m=1}^M$ mean

[9] If $-\log p(y_{1:N}|f) = \sum_{n=1}^N c(y_n, f(x_n))$, we can always take $\ell_N(f(x_{1:N})) = \sum_{n=1}^N c(y_n, f(x_n))$.

[10] Importantly, this does *not* imply that τ_∞ is the limiting measure of (6.12), since the SDE it evolves according to still relies on the (now exact) projection operator defined in (6.13).

that this leaves the approximation $\tau_\infty \approx \widehat{\tau}_\infty$ unaccounted for, the resulting analysis will still be meaningful for the PLS algorithm proposed in Section 6.5.

6.6.2 Characterising Optimal Approximations

In Theorem 6.6.2, we will show that $\widehat{\Pi}$ is close to optimal amongst a class of approximate posteriors. To define this class, we re-interpret τ_∞ in (6.19) as a parameter on $\mathcal{P}_2(\mathbb{R}^M)$ indexing our approximation. Making this explicit, we define for each $\tau \in \mathcal{P}_2(\mathbb{R}^M)$ the measure

$$\widehat{\Pi}_\tau := \int \mathbb{P}\big(F \in \cdot \,|\, F^{1:M} = u\big)\, d\tau(u). \tag{6.20}$$

For each choice of M, the collection of all of these measures can now be understood as a non-parametric variational family given by

$$\mathcal{Q}_M := \Big\{\widehat{\Pi}_\tau : \tau \in \mathcal{P}_2(\mathbb{R}^M)\Big\} \subset \mathcal{P}_2(H_k). \tag{6.21}$$

Notice also that as previously remarked upon in the discussion of (6.17), we know that $\Pi^* \in \mathcal{Q}_N$ (cf. Appendix D.8). Surprisingly, we can give a closed form for the optimal variational approximation to Π^* in $\mathcal{Q}_M$—even when $M < N$ so that $\Pi^* \notin \mathcal{Q}_M$.

Theorem 6.6.1. *For the optimal variational approximation Π_M^* of Π^* over $\mathcal{Q}_M$ given by*

$$\Pi_M^* := \operatorname*{arg\,min}_{Q \in \mathcal{Q}_M} \mathrm{KL}(Q, \Pi^*),$$

we have that $\Pi_M^* = \widehat{\Pi}_{\tau^*}$ with probability measure $\tau^*(u) \propto \exp\left(-V^*(u)\right)$, where

$$V^*(u) := -\log \frac{d\tau^*}{du}(u) = \mathbb{E}_\xi\Big[\ell_N\big(\mu_u(x_{1:N}) + \sqrt{\Sigma(x_{1:N})}\xi\big)\Big] + u^T \frac{1}{2}\Lambda_M^{-1} u \tag{6.22}$$

for all $u \in \mathbb{R}^M$. Here, $\xi \sim \mathcal{N}(0, I_M)$, $\Lambda_M := \mathrm{diag}(\lambda_1, \ldots, \lambda_M)$, and for $u \in \mathbb{R}^M$,

$$\begin{aligned} \mu_u(x_{1:N}) &:= u^T e^{1:M}(x_{1:N}) \in \mathbb{R}^N, \\ \Sigma(x_{1:N}) &:= r(x_{1:N}, x_{1:N}) - e^{1:M}(x_{1:N})^T \Lambda_M e^{1:M}(x_{1:N}) \in \mathbb{R}^{N \times N}, \end{aligned} \tag{6.23}$$

where $e^{1:M}(x_{1:N}) := \big(e^m(x_n)\big)_{m,n=1}^N \in \mathbb{R}^{M \times N}$.

Intriguingly, for the special case of Gaussian likelihood functions, V^* in the above result is quadratic, so that τ^* is a Gaussian measure. This allows us to relate our findings to SVGPs, which variationally approximate GP posteriors using a version of (6.20) that forces τ to be a Gaussian measure on $\mathbb{R}^M$. In light

of this, an implication of Theorem 6.6.1 is that an SVGP using $F^{1:M}$ as its inducing features is optimal *only* under Gaussian likelihoods. In contrast, outside of the Gaussian likelihood setting, the Gaussianity enforced upon τ implies that SVGPs are increasingly poor approximations the more non-Gaussian the measure τ^*.

6.6.3 Optimal Approximations & PLS

Having established the optimal approximation Π_M^* in Theorem 6.6.1, our next result shows that the approximation $\widehat{\Pi}$ targeted by PLS is provably close to Π_M^* under standard regularity assumptions on the negative log likelihood ℓ_N. To do this, we first derive the an explicit form for the limiting measure τ_∞ featuring in $\widehat{\Pi}$ as per (6.19). In particular, we can show (cf. Appendix D.10) that $\tau_\infty(u) \propto \exp\left(-V_\infty(u)\right)$ for all $u \in \mathbb{R}^M$ and

$$V_\infty(u) := \ell_N\big(\mu_u(x_{1:N})\big) + +\frac{1}{2}u^T\Lambda_M^{-1}u,$$

where we use $\mu_u(x_{1:N}) \in \mathbb{R}^N$ as defined in (6.23). To arrive at this result, one simply recognises the SDE in (6.14) as a finite-dimensional Langevin diffusion, and then identifies V_∞ as the potential associated to it.

In a similar vein, V^* defined in Theorem 6.6.1 can be interpreted as the potential of a Langevin diffusion with stationary distribution τ^*. This raises an immediate question: given that τ^* parameterises the optimal variational measure Π_M^*, what stops us from constructing a conventional Langevin sampler based directly on V^*? The answer is simple: V^* depends on an intractable and M-dimensional integral. In contrast, V_∞ is tractable, so that constructing PLS as the Langevin diffusion that draws approximate samples from τ_∞ is computationally feasible.

While V_∞ and V^* are very different in terms of the computational effort required to evaluate them, there is reason to hope that their numerical differences might relatively small. Specifically, a close inspection reveals that V_∞ is a simple approximation to V^*: instead of evaluating the intractable integral by averaging over $\xi \sim \mathcal{N}(0, I_M)$ as in V^*, the potential V_∞ uses a somewhat crude approximation, and evaluates the integrand only at the mode for $\xi = 0$. While crude, this is not a bad approximation strategy for sufficiently large M. In fact, the next result quantifies the resulting error between $\widehat{\Pi}$ and Π_M^* in KL divergence, and shows it to be negligible whenever the eigenvalues $\{\lambda_m\}_{m>M}$ decay sufficiently quickly and the negative log likelihood is both Lipschitz continuous and convex.

Theorem 6.6.2. *Assume that for some $\kappa > 0$, $\ell_N : \mathbb{R}^N \to \mathbb{R}$ is a κ-Lipschitz continuous and convex*

function. Then for any fixed $x_1, \ldots, x_N \in \mathcal{X}$, we have

$$\mathrm{KL}\left(\widehat{\Pi}, \Pi_M^*\right) \leq \frac{\kappa^2}{2} tr\big[\Sigma(x_{1:N})\big] = \frac{\kappa^2}{2} \sum_{m=M+1}^{\infty} \lambda_m \sum_{n=1}^{N} (e_m(x_n))^2. \tag{6.24}$$

Further, if $x_1, \ldots, x_N \in \mathcal{X}$ are independently and identically distributed according to ν, then

$$\mathbb{E}_{x_{1:N}}\Big[\mathrm{KL}\left(\widehat{\Pi}, \Pi_M^*\right)\Big] \leq \frac{N\kappa^2}{2} \sum_{m=M+1}^{\infty} \lambda_m^2. \tag{6.25}$$

As discussed in Section 6.5, the speed of decay in λ_m is typically very fast (cf. Table 6.1). Under Lipschitz continuity and convexity for ℓ_N, the measure $\widehat{\Pi}$ targeted by PLS is therefore guaranteed to be very close to the optimum Π_M^*. Note that these conditions are satisfied in a variety of settings of interest, for instance for the binary classification losses with logistic link functions as in (6.7). For a comprehensive overview of other losses satisfying these conditions, see Steinwart and Christmann [2008].

It is important to note that assuming Lipschitz continuity and convexity for ℓ_N is sufficient, but definitely not necessary. In line with this, we find that empirically, $\widehat{\Pi}$ continues to be an excellent posterior approximation even when these conditions are violated. Lipschitz continuity and convexity should therefore be understood as technical requirements that are likely far too strict. In fact, the negative log Gaussian likelihood $\ell_N(f(x_{1:N})) = \frac{1}{2\sigma^2}||y_{1:N} - f(x_{1:N})||^2$ is a simple example demonstrating that targeting $\widehat{\Pi}$ can be justified without the prerequisite regularity conditions on ℓ_N. Specifically, even though the Gaussian likelihood is not Lipschitz continuous, $\widehat{\Pi}$ is in fact *exactly equal* to Π_M^* (see Lemma D.12.2 in Appendix D.12 for a proof).

Lemma 6.6.3. *If $\ell_N\big(f(x_{1:N})\big) = \frac{1}{2\sigma^2}||y_{1:N} - f(x_{1:N})||^2$ then $\widehat{\Pi} = \Pi_M^*$ is a Gaussian measure.*

Both Theorem 6.6.2 and Lemma 6.6.3 only consider the discrepancy between $\widehat{\Pi}$ and Π_M^*—not that between $\widehat{\Pi}$ and the full Bayes posterior Π^*. For Gaussian likelihoods, this discrepancy has been studied thoroughly in Burt et al. [2019], which showed that for this case, the KL divergence between Π^* and Π_M^* can be upper bounded similarly to (6.24) and (6.25). Since $\widehat{\Pi} = \Pi_M^*$ for this setting by Lemma 6.6.3, the results in Burt et al. [2019] thus transfer to the target measure $\widehat{\Pi}$ of PLS. While similar bounds for the case of non-Gaussian likelihoods are not known, the fact that $M = N$ implies $\Pi^* \in \mathcal{Q}_N$ combined with the typically rapid decay of $\{\lambda_m\}_{m>M}$ should make us hopeful that for large enough M we will have $\Pi_M^* \approx \Pi^*$.

6.6.4 Related Approaches

Continuing the comparison with SVGPs [Titsias, 2009a], Lemma 6.6.3 tells us that the measure $\widehat{\Pi}$ targeted by PLS is not only equal to the optimal variational measure Π_M^*, but also to the SVGP approximation in the case of a Gaussian likelihood.[11] Unlike the SVGP, the measure $\widehat{\Pi}$ targeted by PLS does not force τ in (6.21) to be Gaussian, and is therefore close to the optimal variational measure Π_M^* for non-Gaussian likelihoods, too (cf. Theorem 6.6.2).

Beyond SVGPs, Hensman et al. [2015] propose a Hamiltonian Monte Carlo (HMC) sampler reliant on the sufficiency assumption of SVGP to perform joint posterior inference over kernel hyperparameters and inducing features. Their method can be recast as performing HMC on potential V^* in (6.22), where the integral is approximated using Gauss-hermite quadrature, and where Bayesian inference is extended to the kernel hyperparameters. Importantly, they do not show that the proposed HMC sampler is close to the optimal M-dimensional approximation Π_M^* as we do for PLS in Theorem 6.6.2 and have to propose several heuristics to handle the additional intractabilities.

6.7 Experiments

In this section, we compare PLS to SVGP (with inducing features $U_m := F(z_m)$, $m = 1, \dots, M$). In general, we use the same techniques to choose kernel hyper-parameters and inducing points $z_{1:M}$ for both methods (see Appendix D.14 for details) and use $M = \sqrt{N}$ inducing points. Notice that PLS relies implicitly on inducing points $z_{1:M}$, since the estimation of the M largest eigenvalue-eigenfunction pairs via the Nyström method is based on the kernel matrix $\frac{1}{M} k(z_{1:M}, z_{1:M})$ (cf. Appendix D.7).

Regression Table 6.2a compares SVGP to PLS on various benchmark data sets. Since the regression likelihood is Gaussian, we know both SVGP and PLS target the optimal M-dimensional approximation Π_M^* (cf. Lemma 6.6.3). Unsurprisingly, we obtain similar behaviour, although PLS seems to have a slight edge.

Binary Classification In Table 6.2b we compare PLS and SVGP in a binary classification task. This is a convex and Lipschitz continuous loss and therefore the optimality result in Theorem D.11.1 for PLS applies. SVGP on the other hand makes an error due to the embedded parametric Gaussian assumption. However, since the likelihood is still convex, the optimal M-dimensional posterior will be unimodal and may be close enough to a Gaussian distribution for SVGP to still perform well. Overall, both methods perform similarly.

[11] If the GP is specified via kernel r, and the inducing features are chosen as $U_m = \langle F, e_m \rangle$

Table 6.2: Comparison between PLS and SVGP on various benchmark data sets. We see that SVGP and PLS perform similarly, since in both cases, the negative log-likelihood is a convex function. Performance is averaged over five different train/test splits.

(a) Regression

	PLS	SVGP
	Negative Log Likelihood	
Boston	**0.845 (0.315)**	0.918 (0.162)
Concrete	**1.160 (0.103)**	1.272 (0.084)
Energy (cooling)	**0.724 (0.119)**	1.285 (0.060)
Energy (heating)	0.223 (0.061)	**0.198 (0.056)**
Kin8nm	0.890 (0.057)	**0.778 (0.033)**
Power	**1.321 (0.015)**	1.422 (0.015)
Wine (quality)	**1.333 (0.049)**	1.343 (0.045)
Yacht	1.394 (0.286)	**-0.406 (0.183)**
	Mean Absolute Error	
Boston	**0.361 (0.057)**	0.399 (0.046)
Concrete	**0.603 (0.061)**	0.693 (0.059)
Energy (cooling)	**0.383 (0.034)**	0.774 (0.050)
Energy (heating)	0.229 (0.012)	**0.226 (0.015)**
Kin8nm	**0.407 (0.011)**	0.411 (0.012)
Power	**0.760 (0.016)**	0.872 (0.014)
Wine (quality)	**0.771 (0.036)**	0.782 (0.032)
Yacht	0.256 (0.114)	**0.124 (0.019)**

(b) Binary Classification

	PLS	SVGP
	Area Under the Curve	
Breast	98.37 (0.90)	**98.44 (0.81)**
Diabetes	**83.38 (2.93)**	82.95 (2.59)
Heart	**91.34 (0.75)**	91.21 (0.98)
Ionosphere	90.94 (2.85)	**92.94 (1.71)**
Mushrooms	**81.77 (2.09)**	81.76 (1.95)
Rice	**94.65 (1.54)**	94.04 (1.40)
Wine (colour)	94.95 (1.67)	**94.99 (1.34)**
Yeast	**69.16 (2.39)**	67.23 (1.03)
	Accuracy	
Breast	94.74 (1.48)	**94.74 (0.83)**
Diabetes	**76.56 (4.33)**	76.46 (4.44)
Heart	**83.35 (1.83)**	83.11 (1.41)
Ionosphere	85.91 (4.47)	**87.27 (5.35)**
Mushrooms	**75.64 (1.64)**	74.83 (2.85)
Rice	**88.96 (1.14)**	88.46 (0.56)
Wine (colour)	**93.98 (0.63)**	85.98 (3.34)
Yeast	**64.93 (2.20)**	64.13 (1.30)

Poisson Regression In Figure 6.1 we consider synthetic data for Poisson regression with unknown rate function modelled as f^2. This corresponds to the likelihood $p(y_n|f) = (y_n!)^{-1}(f(x_n))^{2y_n}\exp(-(f(x_n))^2)$, $y_n \in \mathbb{N}_0$, which is non-convex in f due to the symmetry $p(y_n|f) = p(y_n|-f)$. Consequently, the theory in Theorem 6.6.2 does not apply and we do not have guarantees for PLS. However, PLS still handles this situation remarkably well and perfectly models the bimodal nature of the posterior which is reflected by the symmetry around the x-axis. This should be contrasted with SVGP, where no implementation of such a likelihood is available in standard libraries [Gardner et al., 2018] and in any case bimodal functional posteriors are prevented from occurring sine the SVGP posterior is always a Gaussian process.

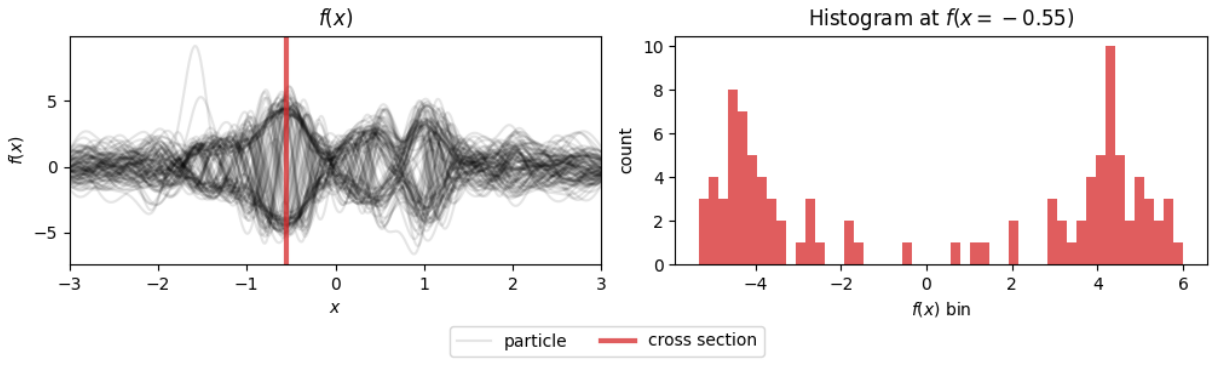


Figure 6.1: This figure illustrates Poisson regression with rate function f^2. We draw $J = 100$ posterior samples from PLS. Due to the non-convexity in the likelihood, we see a bimodal behaviour in function space reflected by a symmetry around the x-axis. The histogram depicts the posterior distribution for $F(x)$ at $x = -0.55$ which is clearly non-Gaussian.

6.8 Conclusion

In this paper we derived the Wasserstein gradient flow for Bayesian inference in function space with a Gaussian random element prior distribution and for arbitrary likelihood functions. We further demonstrated that the WGF can be efficiently implemented by a projected version of the Langevin equation in Hilbert space which is why we call out method projected Langevin sampling (PLS).

PLS coincides with the optimal M-dimensional posterior approximation for Gaussian likelihoods and is the first method to be provably close to the optimal M-dimensional posterior approximation—if Eigenvalues-Eigenfunction paris are known or well approximated—for non-Gaussian likelihoods under convexity and Lipschitz assumptions.

PLS performs competitively in regression and classification tasks on various Benchmark data sets against it's natural competitor SVGP. However, our method is much more widely applicable and we demonstrate its capability of producing multimodal functional posteriors.

Statement of Authorship for joint/multi-authored papers for PGR thesis

Title of Paper	Bayesian Inference in Function Space via the Wasserstein Gradient Flow
Publication Status	Unpublished and unsubmitted work, written in a manuscript style.
Publication Details	None

Student Confirmation

Student Name:	Veit David Wild		
Contribution to the Paper	• Derivation of all theoretical results • Writing of the manuscript • Assistance in implementation of the algorithm		
Signature	V. Wild	Date	28.01.2024

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Dino Sejdinovic		
Supervisor comments			
Signature	Dino Sejdinovic	Date	30 January 2024

This completed form should be included in the thesis, at the end of the relevant chapter.

7 | Conclusion and Future Directions

This dissertation has significantly advanced the field of machine learning and uncertainty quantification in several key areas. Beginning with a concise overview of my contributions, I proceed to assess the extent to which the research objectives were realised. I furthermore discuss limitations, highlight avenues for future research, and conclude with some final remarks.

7.1 Summary

First, a rigorous framework was established for variational inference with Gaussian processes, capitalising on the connection between Gaussian processes and Gaussian random elements in Banach spaces. Subsequently, leveraging this framework, generalised variational inference (GVI) was extended for infinite-dimensional functional parameter spaces. This extension led to the formulation of Gaussian Wasserstein inference, a highly scalable method for function space inference that obtains cutting-edge performance on standardised benchmark datasets.

Second, the challenges associated with parameterisation in GVI spurred the development of an innovative inference method capable of executing gradient descent directly in the space of probability measures. The resulting algorithm interlinks heuristic approaches like deep ensembles with Bayesian methodologies such as the Langevin algorithm. Moreover, it facilitates the derivation of novel GVI procedures like deep repulsive Langevin ensembles, for which we can offer asymptotic convergence guarantees.

Finally, the two main ideas of this thesis were combined: GVI for an infinite-dimensional functional parameter space with a gradient descent performed directly in the space of probability measures. The resulting method implements a sampling procedure in function space that approximates the infinite-dimensional Langevin equation. We show that our approach is provably close to an optimal variational approximation for non-Gaussian likelihoods and optimal for Gaussian likelihoods.

7.2 Future Work and Limitations

The Gaussian random element regression, introduced in Chapter 3, offers avenues for further exploration. In particular, the role of the Banach space E could be analysed further. For example, in Chapter 4 we used the space of square-integrable functions whereas in Chapter 6 we deployed an RKHS. In general, our choice is mostly guided by the desire for a tractable inference procedure. However, from a modelling perspective we are required to parameterise the covariance operators with covariance kernels. In this context, an overview over how to effectively parameterise the covariance operators for GREs with varying

Banach spaces would be valuable.

Gaussian Wasserstein inference and the broader scope of GVI on function spaces, as detailed in Chapter 4, present an avenue for deeper investigation. Successful uncertainty quantification relies on selecting appropriate hyperparameters for the prior kernel. Our approach hinges on employing type-II maximum likelihood (cf. Appendix B.7) to facilitate this selection, effectively performing standard Gaussian process inference on a subset of the available data. A procedure that does not rely on a Bayesian detour would be preferable. Moreover, an exploration of alternative parameterisations for the variational Gaussian measure could offer valuable insights. We have utilised a neural network for the mean and the SVGP kernel for the variational covariance function, but there might exist other viable choices that deliver comparable performance at lower computational costs.

Deep repulsive Langevin ensembles, detailed in Chapter 5, introduce a repulsive effect via the kernel utilised for the maximum-mean discrepancy. The development of kernels tailored to specific applications with desirable properties might be of interest. In particular, it would be interesting to study scenarios, maybe for parameter spaces of moderate size, where there is a clear advantage of DRLE over DE or DLE. Another intriguing avenue involves understanding the bias introduced by Euler–Maruyama discretisation, which might be possible in scenarios where the loss exhibits convexity or other simplifying properties. Exploring this could yield insights into understanding its impact at least in specific situations.

Projected Langevin sampling as introduced in Chapter 6 offers avenues for theoretical investigation. In particular, it would be interesting to investigate how close the optimal variational approximations is to the true posterior for non-Gaussian likelihoods as extension of the results in Burt et al. [2019]. Furthermore, a deeper investigation of the approximation error introduced by the Nyström method might be illuminating. From a methodological point of view, it would be interesting to explore preconditioning techniques for simulation of the Langevin diffusion, since this usually leads to faster convergence and more stable implementations.

7.3 Discussion

At the beginning of my doctoral studies, I was passionate about applying concepts from infinite-dimensional analysis and probability theory to machine learning problems. Naturally, I was driven towards Gaussian processes, which—at least on on a conceptual level—are typically described as “random functions”. Clearly, there should be ample opportunity to apply the advanced mathematical concepts from my graduate studies to problem areas arising in Gaussian process research. However, as passionate as I am about formalism, mathematical consistency and generalisations, I ultimately do not see pursuing them

as ends in themselves. I rather see rigorous mathematics as starting point for an analysis where the correct formalism guides you, provides tools and insights, to eventually solve real-world problems.

Although, one might be tempted to agree with me on the practical usefulness of rigorous advanced mathematics for developing algorithms in real world, the famous David Luenberger had the following to say about one area in particular—functional analysis—which I was very keen on employing:

"Some readers may look with great expectation toward functional analysis, hoping to discover new powerful techniques that will enable them to solve important problems beyond the reach of simpler mathematical analysis. Such hopes are rarely realized in practice. The primary utility of functional analysis for the purposes of this book is its role as a unifying discipline, gathering a number of apparently diverse, specialized mathematical tricks into one or a few general geometric principles." [Luenberger, 1968, page 2]

What a discouraging remark! I did not want to merely organise previous work, I wanted to propose new algorithms that improved upon existing ones. Fortunately, I only encountered his remark towards the end of my doctoral studies allowing me to approach my dissertation with the blissful ignorance of a young researcher.

However, his words contain some prophetic wisdom. Frequently, mathematical concepts beyond calculus and linear algebra "only" serve as unifying principles for an array of diverse approaches and heuristics. In particular, Chapter 3 can be seen in this light. Fundamentally, it was a reformulation of many already existing ideas for variational Gaussian processes in the language of functional analysis. Yet, these reformulations are not devoid of any practical utility and serve several purposes.

Firstly, a new way of looking at the same object may provide us with new ideas of how to interact with it, since what is natural from one vantage point may seem obscure from another. Concretely, it was the language developed in Chapter 3 which makes an approach such as Gaussian Wasserstein inference in Chapter 4 natural and feasible. Once variational inference in infinite dimensions is described mathematically and conceptually almost indistinguishably from the finite dimensional case, it is natural to wonder why we cannot simply parameterise the variational Gaussian measure on the function space in terms of infinite-dimensional parameters such as a variational mean and covariance kernel.

Secondly, any mathematical area has its own techniques and tools which can often be applied with minimal effort, once we identify that an object belongs to a certain class. This becomes quite apparent in Chapter 5. Once one realises that GVI is an optimisation problem defined on the space of *all* probability measures, it opens the door to the rich literature on how to perform gradient descent in such spaces [Ambrosio et al., 2005].

Finally, Chapter 6 illustrates the usefulness of Chapter 3 in both ways. Firstly, whilst it is natural to have the idea of applying the Wasserstein gradient flow to the VI optimisation problem in the context of Gaussian random elements in Hilbert spaces, it is almost inconceivable from a Gaussian process vantage point. Secondly, even if one somehow comes up with a similar thought, it is simply not possible to implement such an idea since it requires a deep understanding of the underlying function space through various concepts from infinite-dimensional analysis such as Fréchet derivatives and integration by parts for Gaussian measures on Hilbert spaces.

To summarise: Although Luenberger is right in pointing out that functional analysis often merely acts as conceptual unifier, he does paint a more pessimistic picture than warranted when it comes to its usefulness in practice. My thesis has repeatedly demonstrated that we can indeed arrive at new and powerful inference algorithms using the additional mathematical structure and intuition unraveled when rigorously formalising a problem.

7.4 Conclusion

This dissertation has developed a theory for variational inference and generalised variational inference in infinite-dimensional function spaces. My work demonstrates that concepts from infinite-dimensional analysis such as Gaussian random elements and their corresponding Gaussian measures can help us to better understand and restructure (generalised) variational inference in function spaces. In particular, I showed that these concepts do not only advance our theoretical understanding, but also guide and assist the development of novel competitive inference algorithms.

My thesis draws attention to the beautiful world of infinite-dimensional analysis and its value for the (generalised) variational function space inference community. I want to encourage researchers in this field that it is worth engaging with these concepts even if one's primary goal is the design of better machine learning algorithms.

In a world where "scale is all you need" and machine learning research is increasingly dominated by "GPU rich" commercial actors, universities will have to focus on theoretically advancing the field by gaining deep insights into the structure of machine learning problems. One way of accomplishing this is by leveraging mathematical concepts from some of the greatest minds of the 20^{th} century such as David Hilbert, Stefan Banach, René Maurice Fréchet, Frigyes Riesz, Arthur Sard, Leonard Gross, Alexander Grothendieck, Vladimir Bogachev and Felix Otto. The mathematics is already there, we just have to use it!

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ben Adlam, Jasper Snoek, and Samuel L Smith. Cold posteriors and aleatoric uncertainty. *arXiv preprint arXiv:2008.00029*, 2020.

R. J. Adler and Robert Taylor. *Random Fields and Geometry*. Springer New York, 2007.

Robert J. Adler. An introduction to continuity, extrema, and related topics for general gaussian processes. *Lecture Notes-Monograph Series*, 12:i–155, 1990.

NU Ahmed and Xinhong Ding. On invariant measures of nonlinear Markov processes. *Journal of Applied Mathematics and Stochastic Analysis*, 6(4):385–406, 1993.

Pierre Alquier. Non-exponentially weighted aggregation: regret bounds for unbounded loss functions. In *International Conference on Machine Learning*, pages 207–218. PMLR, 2021a.

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021b.

Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.

Matias Altamirano, François-Xavier Briol, and Jeremias Knoblauch. Robust and scalable Bayesian online changepoint detection. *arXiv preprint arXiv:2302.04759*, 2023.

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

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50:5–43, 2003.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. *Advances in Neural Information Processing Systems*, 32, 2019.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

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

Viorel Barbu and Michael Röckner. From nonlinear Fokker–Planck equations to solutions of distribution dependent sde. *arXiv preprint arXiv:1808.10706*, 2020.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian bounds based on the Rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444, 2016.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, 2004.

Riddhiman Bhattacharya and Tiefeng Jiang. Fast sampling and inference via preconditioned langevin dynamics. *arXiv preprint arXiv:2310.07542*, 2023.

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5): 1103–1130, 2016.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.

Vladimir Bogachev. *Gaussian Measures*. American Mathematical Society, 1998.

Vladimir Bogachev, Giuseppe Da Prato, and Michael Röckner. Existence and uniqueness of solutions for fokker–planck equations on hilbert spaces. *Journal of Evolution Equations*, 10:487–509, 2010.

Vladimir I Bogachev, Giuseppe Da Prato, and Michael Röckner. Parabolic equations for measures on infinite-dimensional spaces. In *Dokl. Math*, volume 78, pages 544–549, 2008.

Molly Bohanon. Lawyer used chatgpt in court—and cited fake cases. a judge is considering sanctions. *Forbes*, 2023. URL https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/?sh=4c6dd5e37c7f.

Alberto Bressan. Tutorial on the center manifold theorem. *Hyperbolic systems of balance laws*, 1911: 327–344, 2003.

Christopher Brislawn. Traceable integral kernels on countably generated measure spaces. *Pacific Journal of Mathematics*, 150(2):229–240, 1991.

David Burt, Carl Edward Rasmussen, and Mark Van Der Wilk. Rates of convergence for sparse variational Gaussian process regression. In *Proceedings of the 36th International Conference on Machine Learning*, pages 862–871, 2019.

David R Burt, Sebastian W Ober, Adrià Garriga-Alonso, and Mark van der Wilk. Understanding variational inference in function-space. *arXiv preprint arXiv:2011.09421*, 2020a.

David R. Burt, Carl Edward Rasmussen, and Mark van der Wilk. Convergence of sparse variational inference in gaussian processes regression. *Journal of Machine Learning Research*, 21(131):1–63, 2020b. URL http://jmlr.org/papers/v21/19-1015.html.

Krzysztof Chalupka, Christopher KI Williams, and Iain Murray. A framework for evaluating approximation methods for gaussian process regression. *Journal of Machine Learning Research*, 14:333–350, 2013.

Laming Chen, Guoxin Zhang, and Eric Zhou. Fast greedy map inference for determinantal point process to improve recommendation diversity. *Advances in Neural Information Processing Systems*, 31, 2018.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pages 1683–1691. PMLR, 2014.

Ching-An Cheng and Byron Boots. Incremental variational sparse gaussian process regression. *Advances in Neural Information Processing Systems*, 29:4410–4418, 2016.

Ching-An Cheng and Byron Boots. Variational inference for gaussian process models with linear complexity. *arXiv preprint arXiv:1711.10127*, 2017.

Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Mmd-bayes: Robust bayesian estimation via maximum mean discrepancy. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–21. PMLR, 2020.

Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Shunshi Zhang. Analysis of langevin monte carlo from poincare to log-sobolev. In *Conference on Learning Theory*, pages 1–2. PMLR, 2022.

Tzuu-Shuh Chiang, Chii-Ruey Hwang, and Shuenn Jyi Sheu. Diffusion for global optimization in R^n. *SIAM Journal on Control and Optimization*, 25(3):737–753, 1987.

Igor Cialenco, Gregory E Fasshauer, and Qi Ye. Approximation of stochastic partial differential equations by a kernel-based collocation method. *International Journal of Computer Mathematics*, 89(18): 2543–2561, 2012.

Tobias Holck Colding and William P Minicozzi II. Lojasiewicz inequalities and applications. *arXiv preprint arXiv:1402.5087*, 2014.

Lehel Csató and Manfred Opper. Sparse on-line gaussian processes. *Neural computation*, 14(3):641–668, 2002.

Giuseppe Da Prato. *An Introduction to Infinite-Dimensional Analysis*. Springer Berlin Heidelberg, 2006.

Giuseppe Da Prato and Jerzy Zabczyk. *Second order partial differential equations in Hilbert spaces*, volume 293. Cambridge University Press, 2002.

Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic equations in infinite dimensions*. Cambridge university press, 2014.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):651–676, 2017.

Francesco D'Angelo and Vincent Fortuin. Repulsive deep ensembles are Bayesian. *Advances in Neural Information Processing Systems*, 34:3451–3465, 2021.

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

Masoumeh Dashti and Andrew M Stuart. The bayesian approach to inverse problems. *arXiv preprint arXiv:1302.6989*, 2013.

Michael F. Driscoll. The reproducing kernel hilbert space structure of the sample paths of a gaussian process. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 26(4):309–316, 1973.

Nicolas Durrande, James Hensman, Magnus Rattray, and Neil D Lawrence. Detecting periodicities with gaussian processes. *PeerJ Computer Science*, 2:e50, 2016.

Vincent Dutordoir, Nicolas Durrande, and James Hensman. Sparse gaussian processes with spherical harmonic features. In *International Conference on Machine Learning*, pages 2793–2802. PMLR, 2020.

David Duvenaud. The kernel cookbook: Advice on covariance functions. *URL https://www. cs. toronto. edu/duvenaud/cookbook*, 2014.

Donald L Ermak. A computer simulation of charged particles in solution. i. technique and equilibrium properties. *The Journal of Chemical Physics*, 62(10):4189–4196, 1975.

Alessio Figalli and Federico Glaudo. *An invitation to optimal transport, Wasserstein distances, and gradient flows*. European Mathematical Society, 2021.

S Filippi, S Flaxman, D Sejdinovic, and J Cunningham. Bayesian learning of kernel embeddings. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*. Association for Computing Machinery, 2016.

Edwin Fong and Chris C Holmes. Conformal bayesian computation. *Advances in Neural Information Processing Systems*, 34:18268–18279, 2021.

Andrew Foong, David Burt, Yingzhen Li, and Richard Turner. On the expressiveness of approximate inference in bayesian neural networks. *Advances in Neural Information Processing Systems*, 33: 15897–15908, 2020.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Vincent Fortuin, Adrià Garriga-Alonso, Florian Wenzel, Gunnar Rätsch, Richard Turner, Mark van der Wilk, and Laurence Aitchison. Bayesian neural network priors revisited. *arXiv preprint arXiv:2102.06571*, 2021.

Tadayoshi Fushiki. Bootstrap prediction and bayesian prediction under misspecified models. *Bernoulli*, 11(4):747–758, 2005.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. GPytorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. *Advances in neural information processing systems*, 31, 2018.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

Pierre Glaser, Michael Arbel, and Arthur Gretton. Kale flow: A relaxed kl gradient flow for probabilities with disjoint support. *Advances in Neural Information Processing Systems*, 34:8018–8031, 2021.

Tilmann Gneiting. Compactly supported correlation functions. *Journal of Multivariate Analysis*, 83(2): 493–508, 2002.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Peter Grünwald. Safe learning: bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 397–420, 2011.

Peter Grünwald and Thijs Van Ommen. Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 2017.

Benjamin Guedj and John Shawe-Taylor. A primer on pac-Bayesian learning. In *ICML 2019-Thirty-sixth International Conference on Machine Learning*, 2019.

Maxime Haddouche and Benjamin Guedj. Wasserstein PAC-Bayes learning: A bridge between generalisation and optimisation. *arXiv preprint arXiv:2304.07048*, 2023.

Martin Hairer. An introduction to stochastic pdes. *arXiv preprint arXiv:0907.4178*, 2009.

Martin Hairer, Andrew M Stuart, Jochen Voss, and Petter Wiberg. Analysis of spdes arising in path sampling. part i: The gaussian case. *Communications in Mathematical Sciences*, 3(4):587–603, 2005.

Martin Hairer, Andrew M Stuart, and Jochen Voss. Analysis of spdes arising in path sampling part ii: The nonlinear case. *The Annals of Applied Probability*, 2007a.

Martin Hairer, Andrew M Stuart, and Jochen Voss. Analysis of spdes arising in path sampling part ii: The nonlinear case. *The Annals of Applied Probability*, 2007b.

Martin Hairer, Andrew M Stuart, and Jochen Voss. Signal processing problems on function space: Bayesian formulation, stochastic pdes and effective mcmc methods, 2011.

Paul R Halmos. *Measure theory*, volume 18. Springer, 2013.

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

James Hensman, Alexander G Matthews, Maurizio Filippone, and Zoubin Ghahramani. MCMC for variationally sparse Gaussian processes. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

James Hensman, Nicolas Durrande, Arno Solin, et al. Variational fourier features for gaussian processes. *J. Mach. Learn. Res.*, 18(1):5537–5588, 2017.

James Hensman, Nicolas Durrande, and Arno Solin. Variational fourier features for gaussian processes. *Journal of Machine Learning Research*, 18(151):1–52, 2018.

Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993.

Milan Hladnik and Matjaž Omladič. Spectrum of the product of operators. *Proceedings of the American Mathematical Society*, 102(2):300–302, 1988.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

Chris C Holmes and Stephen G Walker. Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503, 2017.

Brian R Hunt, Tim Sauer, and James A Yorke. Prevalence: a translation-invariant "almost every" on infinite-dimensional spaces. *Bulletin of the American mathematical society*, 27(2):217–238, 1992.

Hisham Husain and Jeremias Knoblauch. Adversarial interpretation of Bayesian inference. In *International Conference on Algorithmic Learning Theory*, pages 553–572. PMLR, 2022.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In *International Conference on Artificial Intelligence and Statistics*, pages 703–711. PMLR, 2021.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are Bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR, 2021.

Jack Jewson, Jim Smith, and Chris Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

Carl Jidling, Niklas Wahlström, Adrian Wills, and Thomas B Schön. Linearly constrained gaussian processes. *Advances in Neural Information Processing Systems*, 30, 2017.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Andre G Journel and Charles J Huijbregts. *Mining geostatistics*. The Blackburn Press, 1976.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv:1807.02582*, 2018.

Leonid V Kantorovich. Mathematical methods of organizing and planning production. *Management science*, 6(4):366–422, 1960.

Alexander Kechris. *Classical descriptive set theory*, volume 156. Springer Science & Business Media, 2012.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Mohammad Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable bayesian deep learning by weight-perturbation in adam. In *International Conference on Machine Learning*, pages 2611–2620. PMLR, 2018.

Achim Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

Jeremias Knoblauch. Frequentist consistency of generalized variational inference. *arXiv preprint arXiv:1912.04946*, 2019.

Jeremias Knoblauch. *Optimization-centric generalizations of Bayesian inference*. PhD thesis, University of Warwick, 2021.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Doubly robust Bayesian inference for non-stationary streaming data using β-divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 64–75, 2018.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized variational inference: Three arguments for deriving new posteriors. *arXiv preprint arXiv:1904.02063*, 2019.

Vassili N Kolokoltsov. *Nonlinear Markov processes and kinetic equations*, volume 182. Cambridge University Press, 2010.

Fumiyasu Komaki. On asymptotic properties of predictive distributions. *Biometrika*, 83(2):299–313, 1996.

Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin. Kernel Stein discrepancy descent. In *International Conference on Machine Learning*, pages 5719–5730. PMLR, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronoto, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Alexander Kukush. *Gaussian measures in Hilbert space: construction and properties*. John Wiley & Sons, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Miguel Lázaro-Gredilla and Aníbal R Figueiras-Vidal. Inter-domain gaussian processes for sparse inference using inducing features. In *NIPS*, volume 22, pages 1087–1095. Citeseer, 2009.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257. PMLR, 2016.

Felix Leibfried, Vincent Dutordoir, ST John, and Nicolas Durrande. A tutorial on sparse gaussian processes and variational inference. *arXiv preprint arXiv:2012.13962*, 2020.

Maud Lemercier, Cristopher Salvi, Thomas Cass, Edwin V. Bonilla, Theodoros Damoulas, and Terry Lyons. Siggpde: Scaling sparse gaussian processes on sequential data. In *International Conference on Machine Learning*. PMLR, 2021.

Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Fast dpp sampling for nyström with application to kernel methods. In *International Conference on Machine Learning*, pages 2061–2070. PMLR, 2016.

Yingzhen Li and Yarin Gal. Dropout inference in bayesian neural networks with alpha-divergences. In *International conference on machine learning*, pages 2052–2061. PMLR, 2017.

Yingzhen Li and Richard E Turner. Gradient estimators for implicit models. *arXiv preprint arXiv:1705.07107*, 2017.

Moshe Lichman. UCI machine learning repository, 2013.

Mikhail Lifshits. *Lectures on Gaussian Processes*. Springer Berlin Heidelberg, 2012.

Thomas Milton Liggett. *Continuous time Markov processes: an introduction*, volume 113. American Mathematical Soc., 2010.

Jae Hyun Lim, Nikola B Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, et al. Score-based diffusion models in function space. *arXiv preprint arXiv:2302.07400*, 2023.

Jihao Andreas Lin, Javier Antorán, Shreyas Padhy, David Janz, José Miguel Hernández-Lobato, and Alexander Terenin. Sampling from gaussian process posteriors using stochastic gradient descent, 2024.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. When gaussian process meets big data: A review of scalable gps. *IEEE transactions on neural networks and learning systems*, 31(11):4405–4423, 2020.

Qiang Liu. Stein variational gradient descent as gradient flow. *Advances in neural information processing systems*, 30, 2017.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Zhenxin Liu and Jun Ma. Existence, uniqueness and exponential ergodicity under lyapunov conditions for mckean-vlasov sdes with markovian switching. *Journal of Differential Equations*, 337:138–167, 2022.

Christos Louizos and Max Welling. Multiplicative normalizing flows for variational Bayesian neural networks. In *International Conference on Machine Learning*, pages 2218–2227. PMLR, 2017.

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

David G Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1968.

Milan N. Lukić and Jay H. Beder. Stochastic processes with sample paths in reproducing kernel hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969, 2001.

Chao Ma and José Miguel Hernández-Lobato. Functional variational inference based on stochastic process generators. *Advances in Neural Information Processing Systems*, 34, 2021.

Chao Ma, Yingzhen Li, and José Miguel Hernández-Lobato. Variational implicit processes. In *International Conference on Machine Learning*, pages 4222–4233. PMLR, 2019.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015.

Takuo Matsubara, Jeremias Knoblauch, François-Xavier Briol, and Chris J Oates. Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):997–1022, 2022.

Alexander G de G Matthews, James Hensman, Richard Turner, and Zoubin Ghahramani. On sparse variational methods and the kullback-leibler divergence between stochastic processes. In *Artificial Intelligence and Statistics*, pages 231–239. PMLR, 2016.

Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

Alexander Graeme de Garis Matthews. *Scalable Gaussian process inference using variational methods*. PhD thesis, University of Cambridge, 2017.

David A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999a.

David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170. ACM, 1999b.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *International conference on machine learning*, pages 2391–2400. PMLR, 2017.

Jeffrey W Miller and David B Dunson. Robust bayesian inference via coarsening. *Journal of the American Statistical Association*, 2018.

Jeffrey W. Miller and David B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019.

Yuliya Mishura and Alexander Veretennikov. Existence and uniqueness theorems for solutions of mckean–vlasov stochastic equations. *Theory of Probability and Mathematical Statistics*, 103:59–101, 2020.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10 (1-2):1–141, 2017.

James R Munkres. *Topology*. Pearson Education, 2019.

Cameron Musco and Christopher Musco. Recursive sampling for the nyström method. *arXiv preprint arXiv:1605.07583*, 2016.

Tomoyuki Nakagawa and Shintaro Hashimoto. Robust bayesian inference via γ-divergence. *Communications in Statistics-Theory and Methods*, 49(2):343–360, 2020.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Bernt Øksendal. *Stochastic Differential Equations*. Springer Berlin Heidelberg, 2003.

Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz E Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. Practical deep learning with bayesian principles. *Advances in neural information processing systems*, 32, 2019.

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in partial differential equations*, 2001.

Michela Ottobre, Natesh S Pillai, Frank J Pinski, and Andrew M Stuart. A function space hmc algorithm with second order langevin diffusion limit. *Bernoulli*, 2016.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Emanuel Parzen. An approach to time series analysis. *The Annals of Mathematical Statistics*, pages 951–989, 1961.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

Balram S. Rajput. Gaussian measures on lp spaces, $1 \leq p < \infty$. *Journal of Multivariate Analysis*, 2(4): 382–403, 1972.

Balram S. Rajput and Stamatis Cambanis. Gaussian processes and gaussian measures. *The Annals of Mathematical Statistics*, 43(6):1944–1952, 1972.

RV Ramamoorthi, Karthik Sriram, and Ryan Martin. On posterior concentration in misspecified models. *Bayesian Analysis*, 10(4):759–789, 2015.

KPS Bhaskara Rao and M Bhaskara Rao. *Theory of charges: a study of finitely additive measures*. Academic Press, 1983.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, volume 6. International Conference on Representation Learning, 2018.

Klaus Ritter, Grzegorz W Wasilkowski, and Henryk Woźniakowski. Multivariate integration and approximation for random fields satisfying sacks-ylvisaker conditions. *The Annals of Applied Probability*, pages 518–540, 1995.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

Simon Rodriguez-Santana, Bryan Zaldivar, and Daniel Hernandez-Lobato. Function-space inference with sparse implicit processes. In *International Conference on Machine Learning*, pages 18723–18740. PMLR, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution

image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

H. L. Royden and M. P. Fitzpatrick. *Real Analysis - Fourth Edition*. Pearson, 2010.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *NIPS*, pages 1657–1665, 2015.

Tim GJ Rudner, Zonghao Chen, and Yarin Gal. Rethinking function-space variational inference in bayesian neural networks. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2020.

Adil Salim, Anna Korba, and Giulia Luise. The wasserstein proximal gradient algorithm. *Advances in Neural Information Processing Systems*, 33:12356–12366, 2020.

Hugh Salimbeni, Ching-An Cheng, Byron Boots, and Marc Deisenroth. Orthogonally decoupled variational gaussian processes. *Advances in neural information processing systems*, 31, 2018.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55, 2015.

Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7:87–154, 2017.

Frank Schneider, Lukas Balles, and Philipp Hennig. Deepobs: A deep learning optimizer benchmark suite. *arXiv preprint arXiv:1903.05499*, 2019.

Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

Matthias W Seeger, Christopher KI Williams, and Neil D Lawrence. Fast forward selection to speed up sparse gaussian process regression. In *International Workshop on Artificial Intelligence and Statistics*, pages 254–261. PMLR, 2003.

John Shawe-Taylor and Robert C Williamson. A PAC analysis of a Bayesian estimator. In *Annual Workshop on Computational Learning Theory: Proceedings of the tenth annual conference on Computational learning theory*, volume 6, pages 2–9, 1997.

Jiaxin Shi, Shengyang Sun, and Jun Zhu. A spectral approach to gradient estimation for implicit distributions. In *International Conference on Machine Learning*, pages 4644–4653. PMLR, 2018.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18:1257, 2006.

Edward Snelson and Zoubin Ghahramani. Local and global sparse gaussian process approximations. In *Artificial Intelligence and Statistics*, pages 524–531. PMLR, 2007.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15 (1):1929–1958, 2014.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer New York, 2008.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

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

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational bayesian neural networks. In *International Conference on Learning Representations*, 2018.

Saifuddin Syed, Alexandre Bouchard-Côté, George Deligiannidis, and Arnaud Doucet. Non-reversible parallel tempering: a scalable highly parallel mcmc scheme. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):321–350, 2022.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR, 2009a.

Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In *International conference on machine learning*, pages 1971–1979. PMLR, 2014.

Michalis K Titsias. Variational model selection for sparse gaussian process regression. *Report, University of Manchester, UK*, 2009b.

Ba-Hien Tran, Simone Rossi, Dimitrios Milios, and Maurizio Filippone. All you need is a good functional prior for bayesian deep learning. *arXiv preprint arXiv:2011.12829*, 2020.

Mark Van der Wilk, Carl Edward Rasmussen, and James Hensman. Convolutional gaussian processes. *Advances in Neural Information Processing Systems*, 30, 2017.

Mark van der Wilk, Matthias Bauer, ST John, and James Hensman. Learning invariances using the marginal likelihood. *Advances in Neural Information Processing Systems*, 31, 2018.

JMAM Van Neerven. Stochastic evolution equations. *ISEM lecture notes*, 2008.

A Yu Veretennikov. On ergodic measures for McKean-Vlasov stochastic equations. In *Monte Carlo and Quasi-Monte Carlo Methods 2004*, pages 471–486. Springer Berlin Heidelberg, 2006.

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2003.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Grace Wahba. *Spline models for observational data*. SIAM, 1990.

Ke Wang, Geoff Pleiss, Jacob Gardner, Stephen Tyree, Kilian Q Weinberger, and Andrew Gordon Wilson. Exact gaussian processes on a million data points. *Advances in Neural Information Processing Systems*, 32:14648–14659, 2019.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Jakub Świątkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*, 2020.

Andre Wibisono. Proximal langevin algorithm: Rapid convergence under isoperimetry. *arXiv preprint arXiv:1911.01469*, 2019.

Veit Wild and George Wynne. Variational gaussian processes: A functional analysis view. In *International Conference on Artificial Intelligence and Statistics*, pages 4955–4971. PMLR, 2022.

Veit Wild, Motonobu Kanagawa, and Dino Sejdinovic. Connections and equivalences between the nyström method and sparse variational gaussian processes. *arXiv preprint arXiv:2106.01121*, 2021.

Veit David Wild, Robert Hu, and Dino Sejdinovic. Generalized variational inference in function spaces: Gaussian measures meet Bayesian deep learning. *Advances in Neural Information Processing Systems*, 35:3716–3730, 2022.

Veit David Wild, Sahra Ghalebikesabi, Dino Sejdinovic, and Jeremias Knoblauch. A rigorous link between deep ensembles and (variational) bayesian methods. *Advances in Neural Information Processing Systems*, 2023.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Proceedings of the 14th annual conference on neural information processing systems*, pages 682–688, 2001.

Andrew Wilson and Hannes Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *International conference on machine learning*, pages 1775–1784. PMLR, 2015.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

Andrew Gordon Wilson. The case for Bayesian deep learning. *arXiv preprint arXiv:2001.10995*, 2020.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378. PMLR, 2016.

James Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Deisenroth. Efficiently sampling functions from gaussian process posteriors. In *International Conference on Machine Learning*, pages 10292–10302. PMLR, 2020.

Pei-Shien Wu and Ryan Martin. A comparison of learning rate selection methods in generalized Bayesian inference. *Bayesian Analysis*, 18(1):105–132, 2023.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Caiming Zhang and Yang Lu. Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration*, 23:100224, 2021.

Huaiyu Zhu, Christopher KI Williams, Richard Rohwer, and Michal Morciniec. Gaussian regression and optimal finite dimensional linear models. Technical report, Aston University, 1997.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

Appendices

A | Variational Gaussian Processes - A Functional Analysis View

A.1 Proofs of Section 3.3: Gaussian Random Element Regression

Recall that the posterior measure in Section 3.3 satisfies $P^{F|Y=y}(A) = \int_A \frac{p(y|f)}{p(y)} dP(f)$ for all $A \in \mathcal{B}(E)$.

Theorem A.1.1. *1. For any measurable $h : E \to \mathbb{R}^S$*

$$P^{F|Y=y}\big(h^{-1}(B)\big) = \mathbb{P}(h(F) \in B | Y = y). \tag{A.1}$$

2. The posterior measure $P^{F|Y=y}$ is Gaussian with mean $\widetilde{m}$ satisfying

$$\big(\widetilde{m}, T\big)_E = (m, T)_E + C_{TD}(C_{DD} + \sigma^2 I_N)^{-1} y, \tag{A.2}$$

for any $T \in E^*$ and posterior covariance operator $\widetilde{C} : E^* \to E$ satisfying

$$(\widetilde{C}T, T')_E = (CT, T')_E - C_{TD}(C_{DD} + \sigma^2 I_N)^{-1} C_{DT'}, \tag{A.3}$$

for all $T, T' \in E^*$.

Proof. As $P^{F|Y=Y}$ is a regular version of the conditional probability of $F|Y = y$ it satisfies

$$P^{F|Y=y}(A) = \mathbb{P}(F \in A | Y = y) \tag{A.4}$$

for any fixed $y \in \mathbb{R}^N$, $A \in \mathcal{B}(E)$ [Klenke, 2013, Definition 8.28]. Hence for $A = h^{-1}(B)$ with $h : E \to \mathbb{R}^S$ measurable and $B \in \mathcal{B}(\mathbb{R}^N)$

$$P^{F|Y=y}(h^{-1}(B)) = \mathbb{P}(F \in h^{-1}(B) | Y = y) \tag{A.5}$$

$$= \mathbb{P}(h(F) \in B | Y = y), \tag{A.6}$$

where the last line is using $\{\omega \in \Omega \colon F(\omega) \in h^{-1}(B)\} = \{\omega \in \Omega \colon h(F(\omega)) \in B\}$, this proves the first statement.

For the second statement, we can use equation (A.6) with the choice $h = (T, T') \colon E \to \mathbb{R}^2$ for two arbitrary $T, T' \in E^*$. Recall that the random vector $X : \Omega \to \mathbb{R}^N$ is Gaussian if and only if $\alpha^\top X : \Omega \to \mathbb{R}$ is Gaussian for all $\alpha \in \mathbb{R}^N$. Using this we will show random vector $V := \big((F, T)_E, (F, T')_E, Y\big)$ in

$\mathbb{R}^{N+2}$ is Gaussian. Take any $\alpha \in \mathbb{R}^{N+2}$ and set $\varphi = \sum_{n=1}^{N} \alpha_n D_n + \alpha_{N+1} T + \alpha_{N+2} T' \in E^*$, then

$$\alpha^\top V = \sum_{n=1}^{N+2} \alpha_n V_n = (F, \sum_{n=1}^{N} \alpha_n D_n + \alpha_{N+1} T + \alpha_{N+2} T')_E + \sum_{n=1}^{N} \alpha_n \epsilon_n = (F, \varphi)_E + \sum_{n=1}^{N} \alpha_n \epsilon_n, \tag{A.7}$$

is Gaussian for all $\alpha \in \mathbb{R}^{N+2}$ as it is sum of two independent Gaussian distributions, therefore V is Gaussian.

The mean μ_V is given by

$$\mathbb{E}[\alpha^\top V] = (m, \varphi)_E + 0 = \sum_{n=1}^{N} \alpha_n (m, D_n)_E + \alpha_{N+1}(m, T)_E + \alpha_{N+2}(m, T')_E =: \alpha^\top \mu_V, \tag{A.8}$$

and by the characterising property of the covariance operator C we get the covariance matrix Σ_V

$$\mathrm{Cov}[\alpha^\top V, \alpha^\top V] = (C\varphi, \varphi)_E + \sigma^2 \sum_{n=1}^{N} \alpha_n^2 = \alpha^\top \Sigma_V \alpha, \tag{A.9}$$

where the covariance matrix $\Sigma_V \in \mathbb{R}^{(N+2)\times(N+2)}$ is defined as

$$\Sigma_V \coloneqq \begin{bmatrix} (CT, T)_E & (CT, T')_E & C_{TD} \\ (CT', T)_E & (CT', T')_E & C_{T'D} \\ C_{DT} & C_{DT'} & C_{DD} + \sigma^2 I_N. \end{bmatrix} \tag{A.10}$$

We have showed $V \sim \mathcal{N}(\mu_V, \Sigma_V)$ and so using standard conditioning rules for multivariate Gaussians to condition on the last entry of V it is clear that $h(F)|Y = y$ is Gaussian with the desired mean and covariance. □

A.2 Proofs of Section 3.4: Variational Inference for Gaussian Random Elements

Heavy use is made of the transformation rule for measures [Halmos, 2013, Theorem C]. Let $(E, \mathcal{B}_E, \mu)$ be a measure space and $(W, \mathcal{B}_W)$ a measurable space. If $\mathcal{T} : E \to W$ and $\mathcal{G} : E \to [-\infty, \infty]$ are measurable, then

$$\int (\mathcal{G} \circ \mathcal{T})(x)\, d\mu(x) = \int \mathcal{G}(t)\, d\mu^{\mathcal{T}}(t),$$

where the left hand-side exists if and only if the right hand-side exists. Here again $\mu^{\mathcal{T}}(\cdot) := \mu\big(\mathcal{T}^{-1}(\cdot)\big)$ denotes the image measure of μ induced by $\mathcal{T}$.

Measures in the variational family are Gaussian measures

Recall the definition of a measure in the variational family

$$Q(A) \coloneqq \int_A \left(\frac{dQ^L}{dP^L} \circ L\right)(f)\, dP(f)$$

for all $A \in \mathcal{B}(E)$ with $Q^L = \mathcal{N}(\mu, \Sigma)$ and $P^L = \mathcal{N}((m, L)_E, C_{LL})$.

Theorem A.2.1. *1. For any $A \in \mathcal{B}(E)$*

$$Q(A) = \int_{\mathbb{R}^M} \mathbb{P}\big(F \in A \big| U = u\big) dQ^L(u).$$

2. The measure Q is a Gaussian measure with mean m_Q satisfying

$$(m_Q, T)_E = (m, T)_E + C_{LL}^{-1}\big(\mu - (m, L)_E\big), \tag{A.11}$$

for all $T \in E^*$ and covariance operator C_Q satisfying

$$\big(C_Q T, T'\big)_E = (CT, T')_E + C_{TL} C_{LL}^{-1} (\Sigma - C_{LL}) C_{LL}^{-1} C_{LT'}, \tag{A.12}$$

for all $T, T' \in E^*$.

Proof. Firstly, it suffices to prove the statement for sets of the form $A = T^{-1}(B)$, $B \in \mathcal{B}(\mathbb{R})$, $T \in E^*$, as two measures on a Banach space coincide if and only if they coincide for all sets of this form.

We want to apply the transformation rule for measures. To this end set $\mathcal{T} = (T, L) : E \to \mathbb{R}^{M+1}, \mathcal{T}(F) = (T(F), L(F))$ and $V = T(F)$ and $\pi_L\big((x_1, \ldots, x_{M+1})\big) = (x_2, \ldots, x_{M+1})$ for $x \in \mathbb{R}^{M+1}$. Clearly, $\{\omega \in \Omega \colon T(F(\omega)) \in B\} = \{\omega \in \Omega \colon \mathcal{T}(F(\omega)) \in B \times \mathbb{R}^M\}$ and $\pi_L \circ \mathcal{T} = L$.

From this we use the transformation rule

$$\begin{aligned} Q(\{T \in B\}) &= Q(\{\mathcal{T} \in B \times \mathbb{R}^M\}) \\ &= \int_{\mathcal{T} \in B \times \mathbb{R}^M} \left(\frac{dQ^L}{dP^L} \circ L\right)(f) dP(f) \\ &= \int_{\mathcal{T} \in B \times \mathbb{R}^M} \left(\frac{dQ^L}{dP^L} \circ \pi_L \circ \mathcal{T}\right)(f) dP(f) \end{aligned}$$

$$= \int_{B\times\mathbb{R}^M} \frac{dQ^L}{dP^L}(u)dP^{\mathcal{T}}(u,v)$$
$$= \int_{B\times\mathbb{R}^M} \frac{q(u)}{p(u)}p(u,v)d(u,v),$$

where we denote by $p(u)$ the probability density function (pdf) corresponding to P^L, by $q(u)$ the pdf corresponding to Q^L and $p(u,v)$ the joint pdf corresponding to $P^{\mathcal{T}}$. By $p(u,v) = p(v|U=u)p(u)$ and an application of Fubini

$$\begin{aligned} Q(\{T\in B\}) &= \int_{\mathbb{R}^M} \Big(\int_B p(v|U=u)dv\Big)q(u)du \\ &= \int_{\mathbb{R}^M} \mathbb{P}(V\in B|U=u)dQ^L(u), \end{aligned} \tag{A.13}$$

which proves the claim.

For the second statement we maintain the notation $V = (F,T)_E$. We need to show that (A.13) is Gaussian for any choice of $T\in E^*$. It is well known that the conditional distribution $V|U=u$ can be written as

$$V|(U=u) \overset{\mathcal{D}}{=} (m,T)_E + C_{TL}C_{LL}^{-1}\big(u-(m,L)_E\big) + W =: h(u,W), \tag{A.14}$$

where $\overset{\mathcal{D}}{=}$ means equality in distribution and $W\sim\mathcal{N}\big(0,(CT,T)_E - C_{TL}C_{LL}^{-1}C_{LT}\big)$ independently of U.

Since h is linear in U, we know that $h(U,W)$ is Gaussian for $U\sim\mathcal{N}(\mu,\Sigma)$ and the mean and variance can easily be calculated as

$$\mathbb{E}_Q[h(U,W)] = (m,T)_E + C_{TL}C_{LL}^{-1}\big(\mu-(m,L)_E\big) \tag{A.15}$$
$$\mathrm{Cov}_Q[h(U,W)] = \big(C_QT,T\big)_E = (CT,T)_E + C_{TL}C_{LL}^{-1}(\Sigma-C_{LL})C_{LL}^{-1}C_{LT}. \tag{A.16}$$

In other words Q^T is Gaussian for any $T\in E^*$ and we conclude that Q is a Gaussian measure.

To deduce $(C_QT,T')_E$ for two arbitrary elements $T,T'\in E$ we reduce everything to the one-dimensional case. For $\alpha,\beta\in\mathbb{R}$ let $\varphi = \alpha T+\beta T'$ then since $\varphi\in E^*$ and F is Gaussian

$$(F,\varphi)_E = \alpha(F,T)_E + \beta(F,T')_E,$$

is Gaussian. This proves, by definition, that $(F,T)_E$ and $(F,T')_E$ are jointly Gaussian under Q. The mean and variance of $(F,\varphi)_E$, can be calculated from (A.15) and (A.16). Using standard linear algebra

$$\mathrm{Cov}_Q[(F,T)_E,(F,T')_E] = (CT,T')_E + C_{TL}C_{LL}^{-1}(\Sigma-C_{LL})C_{LL}^{-1}C_{LT'}$$

which shows C_Q is as described in (A.12). □

The Kullback-Leibler divergence is tractable

In this section we show that the Kullback-Leibler divergence between the variational measure Q and $P^{F|Y=y}$ can be re-written in a convenient form. This is well-known for finite dimensional Gaussians and has been done for the process view in Matthews et al. [2016] but we have not seen such a derivation for Gaussian measures in Banach spaces so we include it here for completeness.

First, recall the chain rule for Radon-Nikodym derivatives [Halmos, 2013, Chapter 32]. Let μ, ν and η be σ-finite measures on the same measure space. If $\mu \ll \nu$ and $\nu \ll \eta$, then $\mu \ll \eta$ with Radon-Nikodym derrivative given as

$$\frac{d\mu}{d\eta}(f) = \frac{d\mu}{d\nu}(f)\frac{d\nu}{d\eta}(f)$$

for η-almost every f.

Theorem A.2.2. *1. The Kullback-Leibler divergence satisfies*

$$\begin{aligned} KL(Q, P^{F|Y=y}) &= KL(Q, P) - \mathbb{E}_Q[\log p(y|F)] + \log p(y) \\ &= -\mathcal{L} + \log p(y), \end{aligned}$$

for any $y \in \mathbb{R}^N$.

2. The ELBO is tractable and given as

$$\begin{aligned} \mathcal{L} = \sum_{n=1}^{N} \Bigg(& \log \mathcal{N}\big(y_n | (m, D_n)_E + C_{LL}^{-1}\big(\mu - (m, L)_E\big) C_{LD_n}, \sigma^2\big) \\ & - \frac{1}{2\sigma^2}\Big((CD_n, D_n)_E + C_{D_nL} C_{LL}^{-1} (\Sigma - C_{LL}) C_{LL}^{-1} C_{LD_n} \Big) \Bigg) \\ & - KL\big(\mathcal{N}(\mu, \Sigma), \mathcal{N}((F, m)_E, C_{LL})\big) \end{aligned}$$

3. If the prior mean is zero, then the optimal values used in the variational family $\mathcal{Q}$ for μ and Σ are

$$\begin{aligned} \mu^* &= C_{LL}\big(\sigma^2 C_{LL} + C_{LD} C_{DL}\big)^{-1} C_{LD} y \\ \Sigma^* &= C_{LL}\big(C_{LL} + \frac{1}{\sigma^2} C_{LD} C_{DL}\big)^{-1} C_{LL} \end{aligned}$$

which then leads to the optimal mean and covariance satisfying

$$(m_{Q^*}, T)_E = C_{TL}\big(\sigma^2 C_{LL} + C_{LD}C_{DL}\big)^{-1} C_{LD} y \tag{A.17}$$

$$(C_{Q^*}T, T')_E = (CT, T')_E - C_{TL}C_{LL}^{-1}C_{LT'} + C_{TL}\big(C_{LL} + \frac{1}{\sigma^2}C_{LD}C_{DL}\big)^{-1}\big)C_{LT'}, \tag{A.18}$$

for all $T, T' \in E^*$.

Proof. For the proof of the first statement, by Bayes theorem we know that $P^{F|Y=y}$ is dominated by the prior P with $\frac{dP^{F|Y=y}}{dP}(f) = \frac{p(y|F=f)}{p(y)}$ for any $y \in \mathbb{R}^N$. The reverse statement is also true, that P is dominated by $P^{F|=y}$ for fixed $y \in \mathbb{R}^N$. This is a consequence of $f \mapsto \frac{p(y|F=f)}{p(y)} > 0$ since then for any $A \in \mathcal{B}(E)$ with $P(A) > 0$, $P^{F|Y=y}(A) = \int_A \frac{p(y|F=f)}{p(y)} dP(f) > 0$, which the contrapositive of P being dominated by $P^{F|Y=y}$. The Radon-Nikodym in this situation is given as $\frac{p(y)}{p(y|F=f)}$. Finally, by definition of Q it is dominated by P and $\frac{dQ}{dP}(f) = (\frac{dQ^L}{dP^L} \circ L)(f)$ for any $f \in E$.

The chain-rule for Radon-Nikodym derivatives therefore tells us that Q is dominated by $P^{F|Y=y}$ with

$$\frac{dQ}{dP^{F|Y=y}}(f) = \frac{dQ}{dP}(f)\frac{dP}{dP^{F|Y=y}}(f) = (\frac{dQ^L}{dP^L} \circ L)(f)\frac{p(y)}{p(y|F=f)}.$$

This lets us rewrite the KL divergence as

$$\begin{aligned}
KL(Q, P^{F|Y=y}) &= \int_E \log\left(\frac{dQ}{dP^{F|Y=y}}\right)(f) dQ(f) \\
&= \int_E \log\left((\frac{dQ^L}{dP^L} \circ L)(f)\frac{p(y)}{p(y|F=f)}\right) dQ(f) \\
&= \int_E \log\left((\frac{dQ^L}{dP^L} \circ L)(f)\right) dQ(f) + \int_E \log\left(\frac{p(y)}{p(y|F=f)}\right) dQ(f) \\
&= \int_E \log\left(\frac{dQ^L}{dP^L}(u)\right) dQ^L(u) - \int_E \log p(y|F=f) Q(f) + \log p(y) \\
&= KL(Q^L, P^L) - \mathbb{E}_Q[\log p(y|F)] + \log p(y),
\end{aligned}$$

which proves the first statement.

For the second statement, recall the ELBO is $\mathcal{L} = -KL(Q^L, P^L) + \mathbb{E}_Q[\log(y|F)]$. The KL term is clear, since Q^L and P^L are both Gaussian measures with the required mean and covariance. We therefore investigate the log-likelihood term now. Note that $Y_1, \ldots, Y_N$ are conditionally independent given $F = f$. The expected log-likelihood term therefore factorises as

$$\mathbb{E}_Q[\log p(y|F=f)] = \sum_{n=1}^{N} \mathbb{E}_Q[\log p(y_n|F)].$$

Setting $V = D_n(F) = (F, D_n)_E$ and noting that $p(y_n|F = f)$ depends on F only through D_n, meaning $p(y_n|F = f) = p(y_n|V = D_n(f))$, we see

$$\mathbb{E}_Q[\log p(y_n|F)] = \int_E \log p\big(y_n|V = D_n(f)\big)dQ(f) = \int_E \log p\big(y_n|V = v\big)dQ^{D_n}(v).$$

Note that Q^{D_n} is Gaussian with mean $\mu_V := (m, D_n)_E + C_{LL}^{-1}\big(\mu - (m, L)_E\big)C_{LD_n}$ and variance $\sigma_V^2 := (CD_n, D_n)_E + C_{D_nL}C_{LL}^{-1}(\Sigma - C_{LL})C_{LL}^{-1}C_{LD_n}$ as Q is a Gaussian measure. So using the parametric form of the pdf of a Gaussian

$$\begin{aligned}
\mathbb{E}_Q[\log p(y_n|F)] &= \mathbb{E}_Q\big[-\frac{1}{2\sigma^2}(y_n - V)^2 - \log(\sigma) - \frac{1}{2}\log(2\pi)\big] \\
&= \mathbb{E}_Q\big[-\frac{1}{2\sigma^2}(y_n - \mu_V)^2 - \log(\sigma) - \frac{1}{2}\log(2\pi) - \frac{1}{2\sigma^2}(\mu_V - V_n)^2\big] \\
&= \log \mathcal{N}(y_n|\mu_V, \sigma^2) - \frac{1}{2\sigma^2}\mathbb{E}_Q\big[(\mu_v - V_n)^2\big] \\
&= \log \mathcal{N}(y_n|\mu_V, \sigma^2) - \frac{1}{2\sigma^2}\,\sigma_V^2,
\end{aligned}$$

which proves the claim.

Finally, for the third statement, in Appendix A of Titsias [2009b] the optimal form of μ and Σ are given. Note that we have the same objective function as Titsias [2009b] with the only difference that the kernel matrices k_{nm} and k_{mm} need to be replaced with the covariance matrices C_{LD} and C_{DD}. Plugging in the optimal form for μ^* and Σ^* into (A.11) and (A.12) gives rise to m_{Q^*} and C_{Q^*}. □

A.3 Proof of Section 3.6: Connections between GRE Regression and KRR Nyström

Theorem A.3.1. *Let $F \sim \mathcal{N}(0, C)$ be a GRE in $E = C(\mathcal{X}, \mathbb{R})$ with covariance operator C as defined in (3.6) and assumed pointwise noisy data is observed as described in Section 3.3.2. Let $L_m = \mu_m$, where $\{\mu_m\}_{m=1}^M \subset R(\mathcal{X})$ be the features used in the variational approximation. Set $\mathcal{M} = \{C\mu_m\}_{m=1}^M$ where $C\mu_m = \int k(\cdot, x')d\mu_m(x')$ as the approximating family in the Nyström approximation. Then for $\sigma^2 = N\lambda$ the Nyström KRR estimator $\widehat{f}$ in Section 3.6 is equal to the mean m_{Q^*}, given by (A.17), of the optimal Q^* from the variational family $\mathcal{Q}$.*

Proof. First of all note that $\mathcal{M} \subset H_k$, the RKHS of k. For a proof see [Ghosal and van der Vaart, 2017, Lemma 11.4]. This means that the structure of H_k can be leveraged to deduce $\widehat{f}$. Specifically, as every $f \in \mathcal{M}$ can be expressed as $f = \sum_{m=1}^M \alpha_m C\mu_m$ for some $\alpha \in \mathbb{R}^M$ we can solve the finite dimensional

optimisation problem

$$J(\alpha) \coloneqq \frac{1}{N}\sum_{n=1}^{N}\big(y_n - \sum_{m=1}^{M}\alpha_m C\mu_m(x_n)\big)^2 + \lambda||\sum_{m=1}^{M}\alpha_m C\mu_m||_k^2,$$

in $\alpha \in \mathbb{R}^M$ to find the KRR Nyström estimator. Expanding $J(\alpha)$

$$\begin{aligned} J(\alpha) &= \frac{1}{N}\sum_{n=1}^{N} y_n^2 - 2\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M} y_n\alpha_m g_m(x_n) + \frac{1}{N}\sum_{n=1}^{N}\sum_{m,m'=1}^{M}\alpha_m\alpha_{m'}C\mu_m(x_n)C\mu_{m'}(x_n) \\ &\quad + \lambda\sum_{m,m'=1}^{M}\alpha_m\alpha_{m'}\langle C\mu_m, C\mu_{m'}\rangle_k \\ &= \frac{1}{N}y^\top y - 2\frac{1}{N}y^\top K_{X\mathcal{M}}\alpha + \frac{1}{N}\alpha^\top K_{\mathcal{M}X}K_{X\mathcal{M}}\alpha + \lambda\alpha^\top K_{\mathcal{M}\mathcal{M}}\alpha, \end{aligned}$$

where $(K_{\mathcal{M}X})_{mn} = C\mu_m(x)$ and $(K_{\mathcal{M}\mathcal{M}})_{mm'} = \langle C\mu_m, C\mu_{m'}\rangle_{H_k}$ for $n = 1,\dots,N$ and $m, m' = 1,\dots,M$. Standard rules for differentiation give

$$\begin{aligned} J'(\alpha) &= -\frac{2}{N}K_{X\mathcal{M}}y + \frac{2}{N}K_{X\mathcal{M}}K_{\mathcal{M}X}\alpha + 2\lambda K_{\mathcal{M}\mathcal{M}}\alpha \\ J''(\alpha) &= \frac{2}{N}K_{X\mathcal{M}}K_{\mathcal{M}X} + 2\lambda K_{\mathcal{M}\mathcal{M}}. \end{aligned}$$

It is easy to see that $J'(\alpha) = 0$ for $\alpha = (K_{\mathcal{M}X}K_{X\mathcal{M}} + N\lambda K_{\mathcal{M}\mathcal{M}})^{-1}K_{\mathcal{M}X}y$ and that $J''(\alpha)$ is positive definite. Hence the KRR estimator is given as

$$\widehat{f}(x) = \sum_{m=1}^{M}\alpha_m C\mu_m(x),$$

with $\alpha = (\alpha_1,\dots,\alpha_M)$ given as $\alpha = (K_{\mathcal{M}X}K_{X\mathcal{M}} + N\lambda K_{\mathcal{M}\mathcal{M}})^{-1}K_{\mathcal{M}X}y$.

On the other hand, from (A.17) with $T = \delta_x$ and $D_n = \delta_{x_n}$

$$m_{Q^*}(x) = \sum_{m=1}^{M}\beta_m C\mu_m(x),$$

with $\beta = (\sigma^2 C_{LL} + C_{LD}C_{DL})^{-1}C_{LD}y$.

The only thing left to show is $K_{\mathcal{M}X} = C_{LD}$ and $K_{\mathcal{M}\mathcal{M}} = C_{LL}$. This is a consequence of the relationship between the so-called Cameron-Martin space of a GRE and the RKHS. In particular this exact equivalence is outlined by Ghosal and van der Vaart [2017, Page 316] which completes the proof. □

B | Generalised Variational Inference in Function Space - Gaussian Measures Meet Bayesian Deep Learning

B.1 Bayesian Inference as an Optimisation Problem for an Infinite-Dimensional Prior Measure

Let E be a (infinite dimensional) Polish space and $\mathcal{B}(E)$ the Borel σ-algebra on E. We denote the set of Borel probability measures on $\mathcal{B}(E)$ as $\mathcal{P}(E)$ and choose a fixed prior measure $P \in \mathcal{P}(E)$. The likelihood is described by a Markov kernel function $p : \mathcal{Y} \times E \to [0, \infty)$ with

$$(y, f) \mapsto p(y|f), \tag{B.1}$$

where $\mathcal{Y} \subset \mathbb{R}^N$ is Borel measurable. The prior and the likelihood induce for any fixed $y \in \mathcal{Y}$ a posterior measure denoted as $\widehat{P} \in \mathcal{P}(E)$ [Ghosal and van der Vaart, 2017, Chapter 1.3].

The next theorem shows that this posterior measure is the solution to a certain optimization problem.

Theorem B.1.1 (Bayes posterior as minimiser)**.** *The Bayesian posterior measure $\widehat{P}$ is given as*

$$\widehat{P} = \underset{Q \in \mathcal{P}(E)}{\textit{argmin}} \left\{ -\mathbb{E}_Q\big[\log p(y|F)\big] + \mathbb{D}_{KL}(Q, P) \right\} \tag{B.2}$$

for any fixed prior measure $P \in \mathcal{P}(E)$ and fixed $y \in \mathcal{Y}$ such that $f \in E \mapsto p(y|f) > 0$.

Proof. According to Bayes rule in infinite dimensions [Ghosal and van der Vaart, 2017, Chapter 1.3] we know that $\widehat{P}$ is dominated by P with Radon-Nikodym derivative given as

$$\frac{d\widehat{P}}{dP}(f) = \frac{p(y|f)}{p(y)}, \tag{B.3}$$

for $f \in E$ where $p(y) := \int p(y|F = f)\, dP(f)$ is the marginal likelihood for y. The reverse is also true and P is dominated by $\widehat{P}$. We prove this by contraposition and therefore assume that $P(A) > 0$ for some $A \in \mathcal{B}(E)$. From Bayes rule we know that

$$\widehat{P}(A) = \int_A \frac{p(y|f)}{p(y)}\, dP(f) > 0 \tag{B.4}$$

as the integrand is positive by assumption and $P(A) > 0$. This gives $\widehat{P}(A) > 0$ and therefore that P is

dominated by $\widehat{P}$. In this case standard rules for Radon-Nikodym derivatives give that

$$\frac{dP}{d\widehat{P}}(f) = \frac{p(y)}{p(y|f)}, \tag{B.5}$$

for $f \in E$. Note that without loss of generality we can assume that the optimal $Q \in \mathcal{P}(E)$ is dominated by P (and therefore also dominated by $\widehat{P}$) since otherwise (B.2) is infinite by definition of the KL divergence. For such a Q dominated by P it holds that

$$\begin{aligned} L(Q) &:= -\mathbb{E}_Q\big[\log p(y|F)\big] + \mathbb{D}_{KL}(Q, P) && \text{(B.6)} \\ &= -\int \log p(y|f)\, dQ(f) + \int \log \frac{dQ}{dP}(f)\, dQ(f) && \text{(B.7)} \\ &= -\int \log p(y|f)\, dQ(f) + \int \log \frac{dQ}{d\widehat{P}}(f)\, dQ(f) + \int \log \frac{d\widehat{P}}{dP}(f)\, dQ(f), && \text{(B.8)} \end{aligned}$$

where the last line follows from the chain rule for Radon-Nikodym derivatives. We further have

$$\begin{aligned} L(Q) &= -\int p(y|f)\, dQ(f) + \mathbb{D}_{KL}(Q, \widehat{P}) + \int \frac{p(y|f)}{p(y)}\, dQ(f) \qquad \text{(Bayes Rule)} && \text{(B.9)} \\ &= \mathbb{D}_{KL}(Q, \widehat{P}) + p(y) && \text{(B.10)} \\ &\geq p(y), && \text{(B.11)} \end{aligned}$$

since $\mathbb{D}_{KL}(Q, P) \geq 0$, with equality if and only if $Q = \widehat{P}$. This proves the claim. □

B.2 Pointwise Evaluation as Weak Limit

To outline the problem briefly: If $F \sim \mathcal{N}(m, C)$ is a GRE with mean $m \in L^2(\mathcal{X}, \rho, \mathbb{R})$ and covariance operator C as defined in (4.9) then it is in general unclear what the distribution of $F(x)$ would be for a fixed $x \in \mathcal{X}$. The technical reason is that the pointwise evaluation $\pi_x : L^2(\mathcal{X}, \rho, \mathbb{R}) \to \mathbb{R}$, i.e.

$$\pi_x(f) := f(x) \tag{B.12}$$

is not well-defined. An element g of the space $L^2(\mathcal{X}, \rho, \mathbb{R})$ is an equivalence class and only identifiable up to a ρ-nullset. This means that the definition of π_x in (B.12) makes no sense whenever $\rho(\{x\}) = 0$ which is the case whenever ρ has a pdf w.r.t. the Lebesgue measure.

However, we will remedy this situation by defining for a fixed $x \in \mathcal{X}$

$$F(x) := \lim_{n \to \infty} \langle F, h_{n,x} \rangle_2 \tag{B.13}$$

where $h_{n,x} \in L^2(\mathcal{X}, \rho, \mathbb{R})$ is an appropriately chosen sequence and the limit is to be understood as convergence in distribution of the sequence of scalar random variables $\langle F, h_{n,x}\rangle_2$.

Theorem B.2.1. *Let $F \sim \mathcal{N}(m, C)$ be a GRE in $\mathcal{L}^2(\mathcal{X}, \rho, \mathbb{R})$ with mean $m \in L^2(\mathcal{X}, \rho, \mathbb{R})$ and covariance operator C as defined in (4.9). Assume that ρ is a probability measure on $\mathcal{X} \subset \mathbb{R}^D$ and that ρ is absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}^D$ with pdf ρ'. Denote the support of the measure ρ by $supp(\rho)$ and assume that x is an arbitrary point in the interior of $supp(\rho)$ such that m, k and ρ' are continuous at x.*

Let

$$\eta(t) = \begin{cases} \exp\left(-\frac{1}{1-|t|^2}\right) & \text{if } |t| < 1, \\ 0 & \text{if } |t| \geq 1. \end{cases} \tag{B.14}$$

be the so called standard molifier and note that η is smooth with $\int \eta(t)\, dt = 1$. We further define the sequence $h_{n,x}(t) := \eta\big(n(t-x)\big)/\rho'(t)$ for $n \in \mathbb{N}$, $t \in supp(\rho)$ and $h_{n,x} = 0$ for $t \notin supp(\rho)$. Then

$$\langle F, h_{n,x}\rangle_2 \xrightarrow{\mathcal{D}} \mathcal{N}\big(m(x), k(x,x)\big) \tag{B.15}$$

for $n \to \infty$ where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution.

Proof. Note that $supp(h_{n,x}) = B_{1/n}(x) := \{t \in \mathbb{R}^D : |t - x| \leq \frac{1}{n}\}$ and $B_{1/n}(x) \subset supp(\rho)$ for large enough $n \in \mathbb{N}$ since x is from the interior of $supp(\rho)$. This means that $h_{n,x} \in L^2(\mathcal{X}, \rho, \mathbb{R})$ for large enough n as

$$\int h_{n,x}(t)\, d\rho(t) = \int_{supp(\rho)} \left(\frac{\eta\big(n(t-x)\big)}{\rho'(t)}\right)^2 \rho'(t)\, d\lambda(t) \tag{B.16}$$

$$= \int_{supp(\rho)} \frac{\eta\Big(n(t-x)\Big)}{\rho'(t)}\, dt \tag{B.17}$$

$$= \int_{B_{1/n}(x)} \frac{\eta\Big(n(t-x)\Big)}{\rho'(t)}\, dt. \tag{B.18}$$

The last expression is finite for large enough n because the integrand is continuous at x. According to the definition of of GREs we therefore conclude that

$$\langle F, h_{n,x}\rangle_2 \sim \mathcal{N}\big(\langle m, h_{n,x}\rangle_2, \langle Ch_{n,x}, h_{n,x}\rangle_2\big) \tag{B.19}$$

for large enough $n \in \mathbb{N}$.

The next statement we show is that $m_n(x) := \langle m, h_{n,x}\rangle_2 \to m(x)$ for $n \to \infty$. To this end notice that

$$|m_n(x) - m(x)| = |\int_{B_{1/n}(x)} h_{n,x}(t)\big(m(x) - m(t)\big)\, d\rho(t)| \tag{B.20}$$

$$\leq \int_{B_{1/n}(x)} \eta\Big(n(t-x)\Big)|m(x) - m(t)|\, dt. \tag{B.21}$$

Let now $\epsilon > 0$ be arbitrary. For n large enough we $|m(x) - m(t)| \leq \epsilon$ for all $t \in B_{1/n}(x)$ due to the continuity of m in x. This immediately implies

$$\int_{B_{1/n}(x)} \eta\Big(n(t-x)\Big)|m(x) - m(t)|\, dt \leq \epsilon \int_{B_{1/n}(x)} \eta\Big(n(t-x)\Big)\, dt = \epsilon, \tag{B.22}$$

for large enough n which shows the convergence of $m_n(x)$ to $m(x)$.

A similar argument shows that $k_n(x,x) := \langle Ch_{n,x}, h_{n,x}\rangle_2 \to k(x,x)$ for $n \to \infty$.

We therefore conclude that

$$\langle F, h_{n,x}\rangle_2 = \langle F, h_{n,x}\rangle_2 - m_n(x) + m_n(x) \tag{B.23}$$

$$= \sqrt{k_n(x,x)} \underbrace{\frac{\langle F, h_{n,x}\rangle_2 - m_n(x)}{\sqrt{k_n(x,x)}}}_{\sim\mathcal{N}(0,1)} + m_n(x) \tag{B.24}$$

$$\xrightarrow{\mathcal{D}} \mathcal{N}\big(m(x), k(x,x)\big) \tag{B.25}$$

for $n \to \infty$ due to Slutsky's theorem. □

According to Theorem B.2.1 we can simply define $F(x) \sim \mathcal{N}(m(x), k(x,x))$ for all x in the interior of the support of ρ if m, k and ρ' are continuous at x. These are mild assumptions and we can typically assume that they are satisfied in practice.

B.3 The Wasserstein Metric for Probability Measures

Let E be a Polish space. For $p \geq 1$, let $P_p(E)$ denote the collection of all probability measures μ on E with finite p^{th} moment, that is, there exists some x_0 in M such that:

$$\int_M d(x, x_0)^p \, \mathrm{d}\mu(x) < \infty. \tag{B.26}$$

The p^{th} Wasserstein distance between two probability measures μ and ν in $P_p(E)$ is defined as

$$W_p(\mu,\nu) := \left(\inf_{\gamma\in\Gamma(\mu,\nu)}\int_{E\times E} d(x,y)^p \,\mathrm{d}\gamma(x,y)\right)^{1/p}, \tag{B.27}$$

where $\Gamma(\mu,\nu)$ denotes the collection of all measures on $E\times E$ with marginals μ and ν on the first and second arguments respectively.

More details about the Wasserstein distance can be found in Chapter 7 of Ambrosio et al. [2005].

B.4 A Tractable Approximation of the Wasserstein Metric

Recall that the Wasserstein metric for the two Gaussian measures $P=\mathcal{N}(m_P,C_P)$ and $Q=\mathcal{N}(m_Q,C_Q)$ on the Hilbert space $H=L^2(\mathcal{X},\rho,\mathbb{R})$ is given as

$$W_2^2(P,Q) = ||m_P-m_Q||_2^2 + tr(C_P) + tr(C_Q) - 2\cdot tr\Big[\big(C_P^{1/2}C_QC_P^{1/2}\big)^{1/2}\Big]. \tag{B.28}$$

Further the operators C_P and C_Q are defined through trace-class kernels k and r as described in Section 4.3.1. We will now discuss how to approximate each term in (B.28).

First, note that

$$||m_P-m_Q||_2^2 = \int \big(m_P(x)-m_Q(x)\big)^2\, d\rho(x) \approx \frac{1}{N}\sum_{n=1}^{N}\big(m_P(x_n)-m_Q(x_n)\big)^2, \tag{B.29}$$

which follows by replacing the true input distribution with the empirical data distribution. Second, note that under very general conditions on k and ρ it holds that [Brislawn, 1991]

$$tr(C_P) = \int k(x,x)\, d\rho(x) \tag{B.30}$$

and similarly for C_Q. Again by replacing ρ with the empirical data distribution we obtain natural estimators:

$$tr(C_P) \approx \frac{1}{N}\sum_{n=1}^{N} k(x_n,x_n), \tag{B.31}$$

$$tr(C_Q) \approx \frac{1}{N}\sum_{n=1}^{N} r(x_n,x_n). \tag{B.32}$$

Denote by $\lambda_n(C)$ the n-th eigenvalue of a positive, self-adjoint operator C. By definition of the trace and

the square root of an operator we have

$$tr\Big[\big(C_P^{1/2} C_Q C_P^{1/2}\big)^{1/2}\Big] = \sum_{n=1}^{\infty} \sqrt{\lambda_n\Big(C_P^{1/2} C_Q C_P^{1/2}\Big)} \tag{B.33}$$

$$= \sum_{n=1}^{\infty} \sqrt{\lambda_n\Big(C_Q C_P\Big)}, \tag{B.34}$$

where the second line follows from the fact that the operator $C_Q C_P$ has the same eigenvalues as $C_P^{1/2} C_Q C_P^{1/2}$ [Hladnik and Omladič, 1988, Proposition 1]. The operator $C_Q C_P$ is given as

$$C_Q C_P g(x) = \int r(x, x')(C_P f)(x')\, d\rho(x') \tag{B.35}$$

$$= \int r(x, x')\Big(\int k(x', t) f(t) d\rho(t)\Big)\, d\rho(x') \tag{B.36}$$

$$= \int \int r(x, x') k(x', t) f(t)\, d\rho(x') d\rho(t) \tag{B.37}$$

$$= \int (r * k)(x, t) f(t)\, d\rho(t), \tag{B.38}$$

where we define

$$(r * k)(x, t) := \int r(x, x') k(x', t)\, d\rho(x') \tag{B.39}$$

for all $x, t \in \mathcal{X}$. This means that $C_Q C_P$ is also an integral operator with (non-symmetric) kernel $r * k$. We again replace ρ with $\widehat{\rho}$ to obtain

$$\widehat{(r * k)}(x, t) = \frac{1}{N} \sum_{n=1}^{N} r(x, x_n) k(x_n, t). \tag{B.40}$$

The spectrum of $C_Q C_P$ can now be approximated by the spectrum of the matrix $\frac{1}{N}\widehat{(r * k)}(X, X)$ [Rasmussen and Williams, 2006, cf. Chapter 4.3.2] or $\frac{1}{N_S}\widehat{(r * k)}(X_S, X_S)$ where X_S is a subsample of the data points X of size $N_S < N$. If we plug this approximation into (B.34) we obtain

$$tr\Big[\big(C_P^{1/2} C_Q C_P^{1/2}\big)^{1/2}\Big] \approx \sum_{m=1}^{N_S} \sqrt{\lambda_m\big(\frac{1}{N_S}\widehat{(r * k)}(X_S, X_S)\big)} \tag{B.41}$$

$$= \frac{1}{\sqrt{N_S}} \sum_{m=1}^{N_S} \sqrt{\lambda_m\Big(\frac{1}{N} r(X_S, X) k(X, X_S)\Big)}, \tag{B.42}$$

which is the last expression that we had to approximate.

Note that since $C_Q C_P$ has the same spectrum as the self-adjoint, positive, trace-class operator $C_P^{1/2} C_Q C_P^{1/2}$ we know that its eigenvalues are real, positive and converge to zero.

B.5 Generalised Loss for Regression in Batch Mode

The batch version of the generalised loss is given as:

$$\widehat{\mathcal{L}} = \frac{N}{2}\log(2\pi\sigma^2) + \frac{N}{N_B}\sum_{b=1}^{N_B}\frac{\big(y_{n_b} - m_Q(x_{n_b})\big)^2 + r(x_{n_b}, x_{n_b})}{2\sigma^2} + \frac{1}{N_B}\sum_{b=1}^{N_B}\big(m_P(x_{n_b}) - m_Q(x_{n_b})\big)^2 \tag{B.43}$$

$$+ \frac{1}{N_B}\sum_{b=1}^{N_B} k(x_{n_b}, x_{n_b}) + \frac{1}{N_B}\sum_{b=1}^{N_B} r(x_{n_b}, x_{n_b}) - \frac{2}{\sqrt{N_B N_S}}\sum_{s=1}^{N_S}\sqrt{\lambda_s\big(r(X_S, X_B)k(X_B, X_S)\big)}, \tag{B.44}$$

$N_B \in \mathbb{N}$ is the batch-size. The indices $n_1, \ldots, n_{N_B}$ are the batch-indices and X_B is the batch matrix.

B.6 GWI for (Multiclass) Classification

Let $\{(x_n, y_n)\}_{n=1}^N \subset \mathcal{X} \times \mathcal{Y}$ be data with $\mathcal{X} \subset \mathbb{R}^D$ and $\mathcal{Y} = \{1, \ldots, J\}$, where $J \in \mathbb{N}$ represents $J \geq 2$ distinct classes.

Model We use the same likelihood for $y := (y_1, \ldots, y_N)$ as described in Chapter 4 of Matthews [2017] which is:

$$p(y|f_1, \ldots, f_J) = \prod_{n=1}^{N} p(y_n|f_1, \ldots, f_J) \tag{B.45}$$

with

$$p(y_n|f_1, \ldots, f_J) := h^\epsilon_{y_n}\big(f_1(x_n), \ldots, f_J(x_n)\big), \tag{B.46}$$

for $y_n \in \{1, \ldots, J\}$. The function h^ϵ_ℓ is defined as

$$h^\epsilon_\ell(t_1, \ldots, t_J) \begin{cases} 1 - \epsilon & \text{if } \ell = \underset{j=1,\ldots,J}{\operatorname{argmax}}\{t_j\}, \\ \frac{\epsilon}{J-1} & \text{if otherwise.} \end{cases} \tag{B.47}$$

for $\ell = 1, \ldots, J$ for $\epsilon > 0$. We chose $\epsilon = 1\%$ in our implementation.

We assume that $F_1, \ldots F_J$ are independent GREs on $L^2(\mathcal{X}, \rho, \mathbb{R})$ with prior means $m_{P,j}$ and prior covariance operators $C_{P,j}$, $j = 1, \ldots, J$.

The variational measures for $F_1, \ldots, F_J$ are assumed to be independent and given as $Q_j = \mathcal{N}\big(m_{Q,j}, C_{Q,j}\big)$ for $j = 1, \ldots, J$. We further write $\mathbb{Q}\Big(\big(F_1(x), \ldots, F_J(x)\big) \in A\Big)$, $A \subset \mathbb{R}^J$ for the variational (posterior) approximation of the probability of the event $\{\big(F_1(x), \ldots, F_J(x)\big) \in A\}$.

This leads to the following expected log-likelihood

$$\mathbb{E}_{\mathbb{Q}}\big[\log p(y|F_1,\dots,F_J)\big] \tag{B.48}$$

$$= \sum_{n=1}^{N} \mathbb{E}_{\mathbb{Q}}\big[\log p(y_n|F_1,\dots,F_J)\big] \tag{B.49}$$

$$= \sum_{n=1}^{N} \log(1-\epsilon)\mathbb{Q}\big(\operatorname*{argmax}_{j=1,\dots,J}\{F_j(x_n)\} = y_n\big) + \log(\frac{\epsilon}{J-1})\mathbb{Q}\big(\operatorname*{argmax}_{j=1,\dots,J}\{F_j(x_n)\} \neq y_n\big) \tag{B.50}$$

$$\approx \sum_{n=1}^{N} \log(1-\epsilon)S(x_n,y_n) + \log(\frac{\epsilon}{J-1})\big(1-S(x_n,y_n)\big), \tag{B.51}$$

with

$$S(x,j) := \frac{1}{\sqrt{\pi}} \sum_{i=1}^{I} w_i \prod_{l\neq j} \phi\Big(\frac{\sqrt{2r_j(x,x)}\xi_i + m_{Q,j}(x) - m_{Q,l}(x)}{\sqrt{r_l(x,x)}}\Big) \tag{B.52}$$

for any $x \in \mathcal{X}$, $j = 1,\dots,J$ where $(w_i,\xi_i)_{i=1}^{I}$ are the weights and roots of the Hermite polynomial of order $I \in \mathbb{N}$. This is the same Gauss-Hermite approximation as described in Chapter 4 of Matthews [2017].

The final objective for multiclass classification is given as

$$\mathcal{L} = -\mathbb{E}_{Q}\big[\log p(y|F_1,\dots,F_J)\big] + \sum_{j=1}^{J} W_2^2(P_j,Q_j), \tag{B.53}$$

where the expected log-likelihood is approximated by (B.51) and each Wasserstein distance $W_2^2(P_j,Q_j)$ can be estimated as in (4.14)-(4.15).

Prediction The probability that an unseen point $x^* \in \mathcal{X}$ belongs to class $j \in \{1,\dots,J\}$ is given as

$$\mathbb{Q}(Y^* = j) = (1-\epsilon)S(x^*,j) + \frac{\epsilon}{J-1}\big(1-S(x^*,j)\big) \tag{B.54}$$

for any $x^* \in \mathcal{X}$. We predict the class label as maximiser of this probability. If we apply tempering, we simply replace every $r_j(x,x)$ with $T \cdot r_j(x,x)$ for $j = 1,\dots,J$ in the definition of $S(x,j)$.

Negative Log Likelihood The variational approximation to the negative log-likelihood is

$$NLL = -\log\Big[(1-\epsilon)S(x^*,y^*) + \frac{\epsilon}{J-1}\big(1-S(x^*,y^*)\big)\Big] \tag{B.55}$$

for any point $x^* \in \mathcal{X}$ for which we know that the class label is $y^* \in \{1,\dots,J\}$.

B.7 Implementation Details: Regression

The Regression model is given as $F \sim \mathcal{N}(0, C)$ and

$$Y_n = F(x_n) + \epsilon_n \tag{B.56}$$

with $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$, $n = 1, \dots, N$. The covariance operator C_P depends on the choice of a kernel k, i.e. $C_P = C_{P,k}$ for which we use the ARD kernel k given as

$$k(x, x') = \sigma_f^2 \exp\Big(-\frac{1}{2} \sum_{d=1}^{D} \frac{(x_d - x'_d)^2}{\alpha_d^2} \Big) \tag{B.57}$$

for $x, x' \in \mathbb{R}^D$. We refer to $\sigma_f > 0$ as *kernel scaling factor*, to $\alpha_d > 0$ as *length-scale* for dimension d and to $\sigma > 0$ as *observation noise*.

The data is first randomly split into three categories: training set 80%, validation set 10% and test set 10%. The observations Y are then standardised by subtracting the empirical mean (of the training data) and dividing by the empirical standard deviation (of the training data). The inputs data X is left unaltered.

The number of inducing points The number of inducing points M is treated as a hyperparameter, this means we train the model for each $M \in \{0.5\sqrt{N}, \sqrt{N}, 1.5\sqrt{N}, 2\sqrt{N}\}$ and choose the best model. For GWI: SVGP we use $M \in \{1\sqrt{N}, 2\sqrt{N}, \dots 5\sqrt{N}\}$.

The choice of inducing points The input points $Z_1, \dots, Z_M$ in (4.18) are sampled independently from the training data X and then fixed for GWI-net. For GWI: SVGP they are only initialised this way and then learned by maximising the generalized loss.

Prior hyperparameters The prior hyperparameters σ_f, $\alpha := (\alpha_1, \dots, \alpha_D)$ and σ are chosen by maximising the marginal log-likelihood for the data $X = Z$ and the corresponding observations, which we denote Y_Z. Note that the marginal log-likelihood is tractable and given as

$$\log p(y_Z) = -\frac{1}{2} \log\Big(\det\big(k(Z,Z) + \sigma^2 I_M\big)\Big) - \frac{1}{2} {y_Z}^T \big(k(Z,Z) + \sigma^2 I_M\big)^{-1} y_Z. \tag{B.58}$$

and can therefore be evaluated in $\mathcal{O}(M^3) = \mathcal{O}(N\sqrt{N})$.

Variational mean For GWI-net we use a neural network with $L = 2$ hidden layers, width $D_1 = D_2 = 10$ and tanh as activation function. This follows the set-up of Ma and Hernández-Lobato [2021].

Variational kernel The kernel r which is chosen as described in (4.18) and therefore depends on the covariance matrix $\Sigma \in \mathbb{R}^{M \times M}$ and the $M \in \mathbb{N}$ inducing points $Z = (Z_1, \dots, Z_M) \in \mathbb{R}^{D \times M}$. We

parameterise Σ as $\Sigma = LL^T$ with initialisation

$$L = \text{Chol}\Big(\big(k(Z,Z) + \frac{1}{\sigma^2}k(Z,X)k(X,Z)\big)^{-1}\Big), \tag{B.59}$$

where $k(Z,X)k(X,Z)$ is approximated by batch-sizing as $\frac{N}{N_B}k(Z,X_B)k(X_B,Z)$. This corresponds to an approximation of the optimal choice for Σ in SVGP [Titsias, 2009a].

Parameters in the generalised loss The generalised loss in Appendix B.5 depends further on N_S, N_B and X_S. The batch-size N_B is chosen to be $N_B = 1000$ for $N > 1000$. For $N < 1000$ we use the full training data. The comparison points X_S are sampled independently from the training data X in each iteration. We train here for 1000 epochs on the regression task and 100 epochs on the classification task following Ma and Hernández-Lobato [2021].

Tempering the predictive posterior

Wenzel et al. [2020] observe that the performance of many Bayesian neural networks can be improved by *tempering* the predictive posterior. Tempering refers to a shrinking of the predictive posterior variance by a factor of $\alpha_T \in [0,1]$. This effect has also been observed for Gaussian processes in Adlam et al. [2020] where it can be interpreted as elevating problems that occur from prior misspecification. The prior hyperparameters for the ARD kernel k in (4.16) are selected by maximising the marginal log-likelihood on a subset of the training data. This procedure may lead to prior misspecification, which is why we decided to temper the predictive posterior, which means that we use the predictive distribution

$$Y^*|Y \sim \mathcal{N}\Big(m_Q(x^*), \alpha_T\big(r(x^*,x^*) + \sigma^2\big)\Big) \tag{B.60}$$

for an unseen data point $x^* \in \mathcal{X}$. The (tempered) NLL for each data point is given as

$$\begin{aligned} \text{NLL} &:= -\log p_{\alpha_T}(y^*|y) && \text{(B.61)} \\ &= \frac{1}{2}\log\Big(\alpha_T \cdot (r(x^*,x^*) + \sigma^2)\Big) + \frac{1}{2}\frac{(y-y^*)^2}{\alpha_T \cdot (r(x^*,x^*) + \sigma^2)} + \frac{1}{2}\log(2\pi). && \text{(B.62)} \end{aligned}$$

The tempering factor α_T is chosen as minimiser of the average NLL on the validation set. The final predictions on the test set are made using this optimal α_T and (B.60). Note however that for the NLL numbers reported in Table 4.1 we add $\log(\widehat{\sigma}_{train})$ to (B.62) where $\widehat{\sigma}_{train}$ is the empirical standard deviation of the training data. This is done for fair comparison as it is how the NLL is calculated in Ma and Hernández-Lobato [2021].

B.8 Implementation Details: Classification

As described in section (B.6) we use the prior mean functions $m_{P,j}$ and kernels k_j for $j = 1, \dots, J$. For our experiments we chose $m_{P,j} = 0$ for $j = 1, \dots, J$ and $k := k_1 = \dots, k_J$ where k is the ARD kernel in (4.16).

We use a multi-output neural network for the variational means $m_{Q,j}$ and an SVGP kernel for each r_j , $j = 1, \dots, J$.

The number of inducing points The number of inducing points M is treated as a hyperparameter, this means we train the model for each $M \in \{0.5\sqrt{N}, 0.75\sqrt{N}, \sqrt{N}\}$ and choose the best model.

The choice of inducing points The input points $Z_1, \dots, Z_M$ in (4.18) are sampled independently from the training data X and then fixed for GWI-net.

Prior hyperparameters The prior hyperparameters are initialised as described in B.7, thus maximising the marginal likelihood of a *regression* model, since the marginal likelihood of our classification model is intractable.

Variational mean We use the same CNN architecture as described in Immer et al. [2021], Schneider et al. [2019] for all models.

Variational kernel Each variational kernel r_j uses the same inducing points Z but gets an individual matrix $\Sigma^j \in \mathbb{R}^{M \times M}$ for $j = 1, \dots, J$. They are all initialised as described in B.7.

Parameters in the generalised loss The generalised loss in Appendix B.5 depends on N_S, N_B and X_S. The batch-size N_B is chosen to be $N_B = 1000$ for $N > 1000$. For $N < 1000$ we use the full training data. The comparison points X_S are sampled independently from the training data X in each iteration. We train 100 epochs on the classification task following Ma and Hernández-Lobato [2021].

Tempering the predictive posterior For the same reasons as outlined in Appendix B.7 we temper the predictive posterior. Recall that the NLL for classification is given as

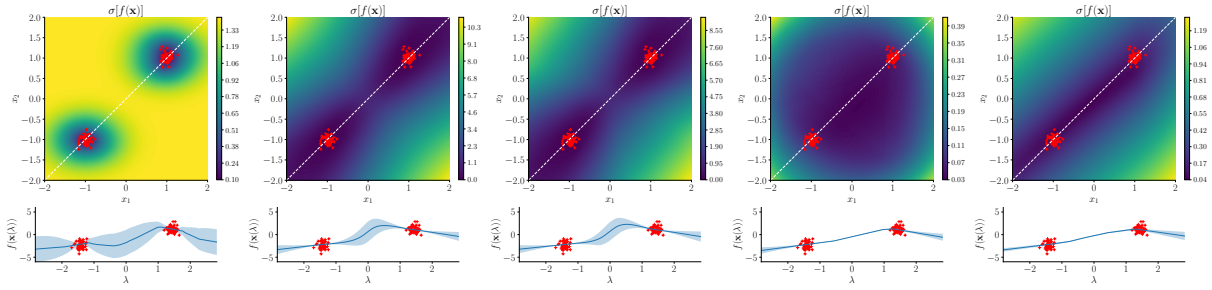
$$NLL = -\log\left[(1-\epsilon)S(x^*, y^*) + \frac{\epsilon}{J-1}\big(1 - S(x^*, y^*)\big)\right] \tag{B.63}$$

for any point $x^* \in \mathcal{X}$ for which we know that the class label is $y^* \in \{1, \dots, J\}$. We use a tempering factor $\alpha_j > 0$ for each variational measure $Q_j \sim \mathcal{N}(m_{Q,j}, \alpha_j r_j)$, $j = 1, \dots, J$. We train the model with $\alpha_j = 1$ for all $j = 1, \dots, J$ and select the tempering factors afterwards as minimiser of the average NLL on the validation set.

B.9 Illustrative Example for Two-Dimensional Inputs

In Foong et al. [2020] it is observed that several BNN posterior approximation techniques struggle with the quantification of in-between uncertainty. The red points mark where observations were made and it is clear that mean-field variational inference (MVFI) [Hinton and Van Camp, 1993] and Monte Carlo Dropout (MCDO) [Gal and Ghahramani, 2016] exhibit unjustifiably high posterior certainty in the area where no observations are made. This is a pathology of the approximation technique as the true Bayesian posterior which is approximated to very high precision by Hamiltonian Monte Carlo (HMC) [Neal, 2012] or the infinite-width GP limit [Matthews et al., 2018] do not display such behaviour.

In Figure B.1 our method GWI-net is displayed next to the methods described in Foong et al. [2020]. As one can observe our model is keenly aware of its limited ability to predict points in-between the two clusters of observed data points.



(a) GWI (b) Inf-width limit GP (c) HMC (d) MFVI (e) MCDO

Figure B.1: Regression on a 2D synthetic dataset (red crosses). The colour plots show the standard deviation of the output, $\sigma[f(\mathbf{x})]$, in 2D input space. The plots beneath show the mean with 2-standard deviation bars along the dashed white line (parameterised by λ). MFVI and MCDO are overconfident for $\lambda \in [-1, 1]$.

m

B.10 Model Misspecification in Gaussian Wasserstein Inference

The generalised loss in Appendix B.5 is a valid optimisation target for any $m_P, m_Q \in L^2(\mathcal{X}, \rho, \mathbb{R})$ and any trace-class kernels k and r. This gives the user a lot of abilities to specify different models, by experimenting with various choices, specifically for m_Q and r. However with great power comes great responsibility: it is quite easy to misspecify GWI. To illustrate the issue let us use a periodic kernel k [Duvenaud, 2014] given as

$$k(x, x') := \sigma_f^2 \exp\big(-\frac{1}{\alpha^2}\sin^2(\pi|x - x'|/p)\big) \tag{B.64}$$

and the SVGP kernel r in (4.18). By the definition of r the uncertainty will be low for points *similiar* to the inducing points Z, i.e. for points $x \in \mathcal{X}$ $k(x, z_m) \approx \sigma_f^2$ for all $m = 1, \ldots, M$. A problem now occurs, if the posterior mean m_Q does not respect the knowledge embedded in k and r. Lets for example use a simple fully connected deep neural network m_Q and choose the point $x^* := z_1 + 10p$. Assume further that $z_1, \ldots, z_M < x^*$. Then we get $k(x^*, z_m) = k(z_1, z_m)$ for all $m = 1, \ldots, M$ due to the periodicity of $\sin(x)$ and therefore $r(x^*, x^*) = r(z_1, z_1)$. It is however very unlikely that the neural network will predict $m_Q(z_1)$ as well as $m_Q(x^*)$ since it is unaware of this periodicity.

This small example should illustrate that it is crucial that m_Q is compatible with the prior knowledge reflected in k and r. However, note that this problem is not present for our model, GWI-net. The ARD kernel encodes the inductive bias that the underlying function is infinitely differentiable and that points close to each other have highly correlated functional outputs. A simple fully connected DNN with tanh activation function is indeed smooth and further it is reasonable to assume that predictions are more unreliable the further they are from the data (as measured by the squared euclidean distance). The ARD kernel is in this sense compatible with a fully connected DNN.

It shall be noted that the DNN used for the classification examples in (4.5) used convolutional layers as explained in Appendix B.8. This can be understood as embedding prior knowledge about translation equivariance into the DNN [Goodfellow et al., 2016, Chapter 9.4]. It might therefore be desirable to use a prior kernel k that embeds similar properties such as the kernel suggested by Van der Wilk et al. [2017]. We considered this to be beyond the scope of this paper but the interaction of DNN architecture and the choice of prior kernels is an interesting avenue for future research.

B.11 Details on Computational Resources

For all our experiments, we distributed our jobs across 8 Nvidia V100 cards.

B.12 Additional Plots for 1D Experiments

In Figure B.2 we compare GWI-net, GWI-SVGP and SVGP on one-dimensional toy data. Note that all three methods use the same posterior kernel, but GWI-net differs from GWI-SVGP in terms of the posterior mean function. GWI-SVGP and SVGP have the same posterior mean but differ in terms of the objective function used for training.

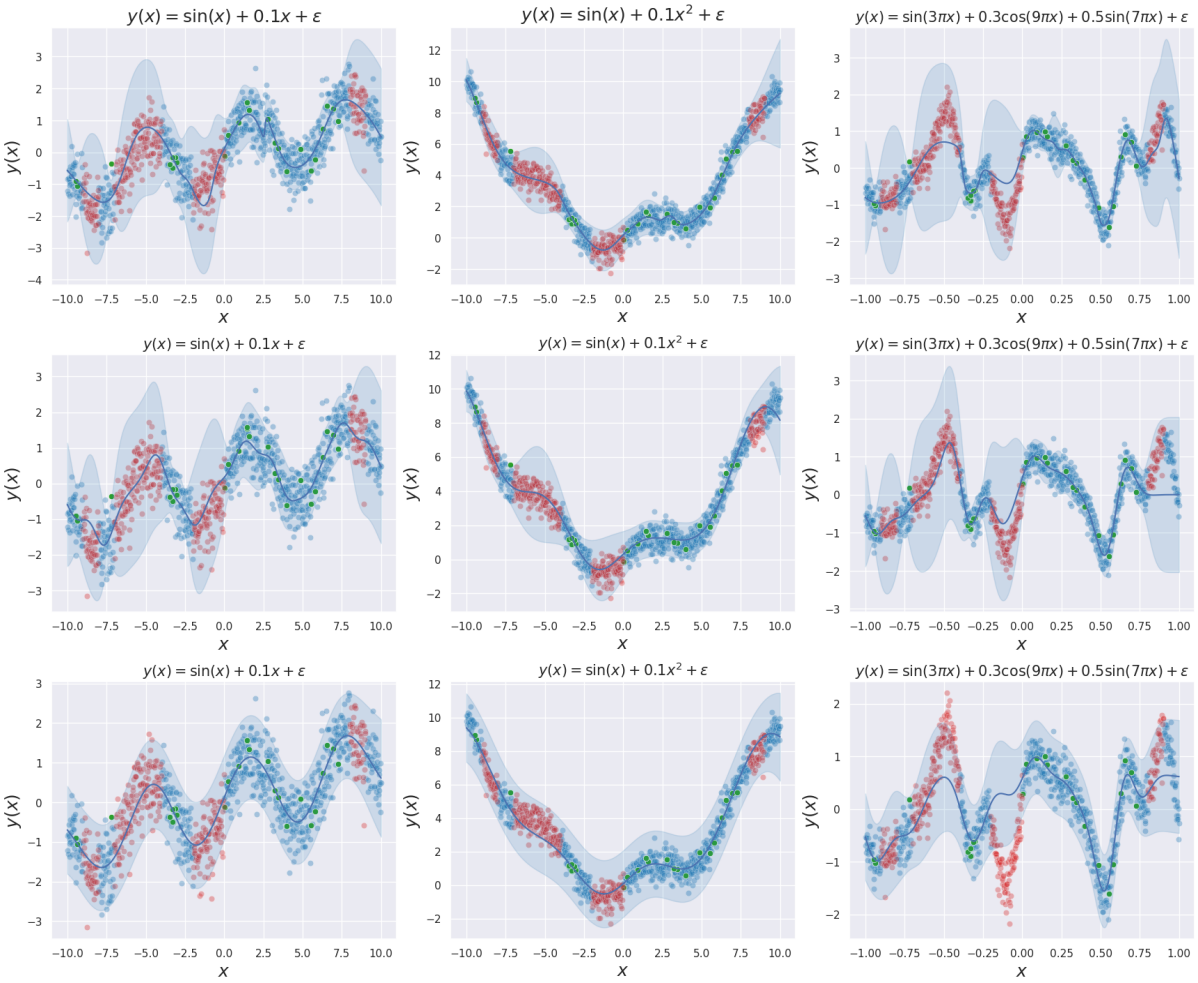
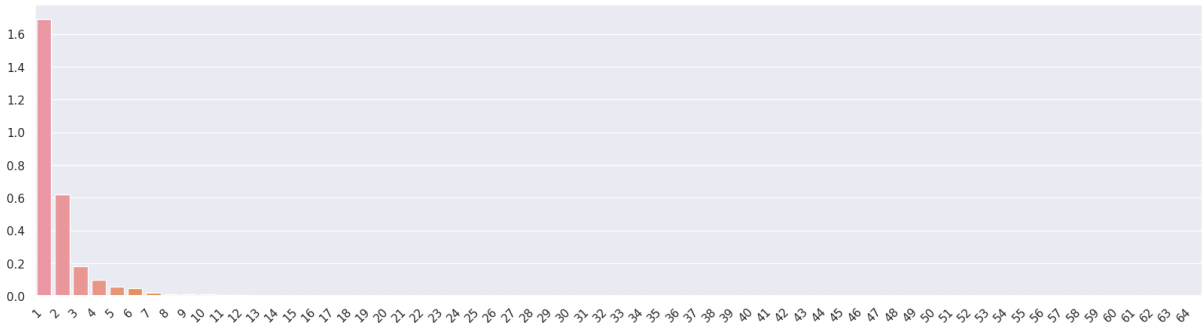


Figure B.2: ■ : Training data ■ : Unseen data ■ : Inducing points
We query the above functions at $N = 1000$ equidistant points and add white noise with $\epsilon \sim \mathcal{N}(0, 0.5^2)$. We use $M = 30$ inducing points and train our method as described in Appendix B.7. The plot shows $m_Q(x) \pm 1.96\sqrt{\mathbb{V}[Y^*(x)|Y]}$ where $\mathbb{V}[Y^*(x)|Y]$ is the posterior predictive variance given as $r(x,x)+\sigma^2$. Here the fitted models from top to bottom are GWI-net, GWI-SVGP and SVGP.

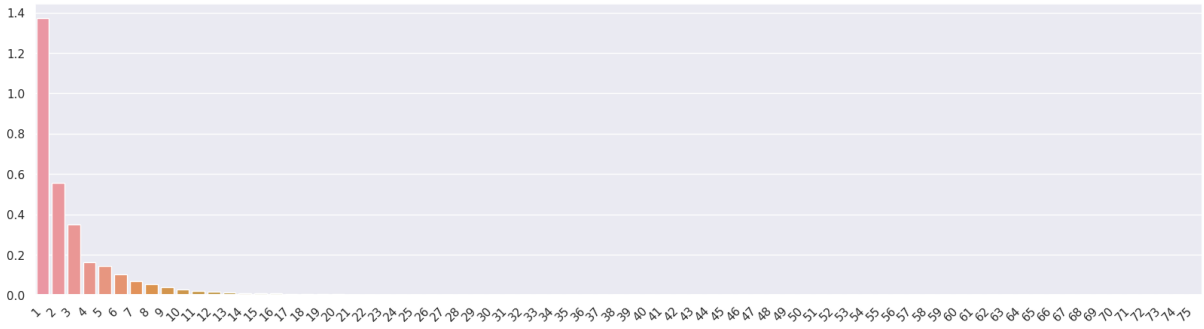
B.13 Empirical estimation error of 2-Wasserstein Distance

The approximation quality of the 2-Wasserstein distance is determined by the approximation quality of the spectrum of the appearing covariance operators. For most kernels in practice like SE or Matern kernel, the spectrum decays very quickly, which is why using the first 100 eigenvalues often empirically seems to be sufficient to approximate the spectrum and therefore the 2-Wasserstein distance. We plot the magnitude of the first 100 positive eigenvalues (sorted on magnitude) for datasets BOSTON, CONCRETE, ENERGY, WINE and YACHT in Figure B.3.

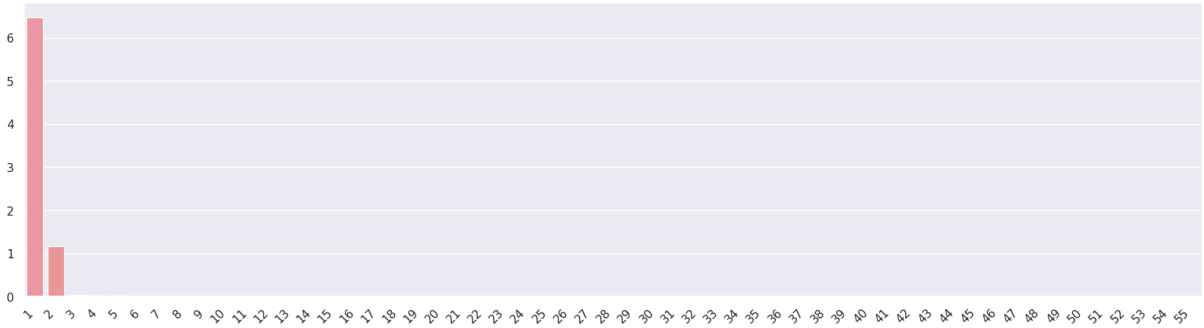
We see in Figure B.3 that eigenvalues indeed decay fast.



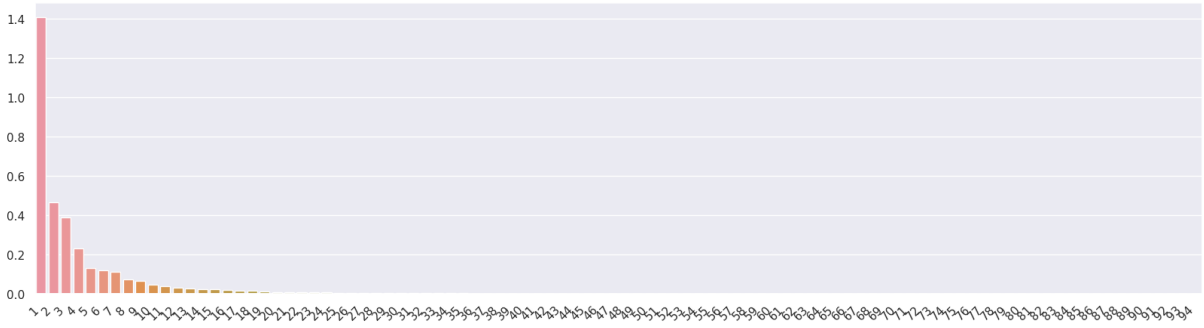
(a) BOSTON



(b) CONCRETE



(c) ENERGY



(d) WINE

Figure B.3: The first 100 positive eigenvalues of $r(X_S, X)k(X, X_S)$ for datasets BOSTON, CONCRETE, ENERGY and WINE

C | A Rigorous Link between Deep Ensembles and Variational Bayesian Methods

C.1 Existence and Uniqueness of Global Minimiser

In this section, we discuss assumptions under which the global minimiser of the optimisation problem

$$L(Q) = \int \ell(\theta)\, dQ(\theta) + \lambda D(Q, P) \tag{C.1}$$

over $\mathcal{P}(\mathbb{R}^J)$ exists and is unique. We assume throughout that the optimisation problem is not pathological, in the sense that there exists a measure $\widehat{Q} \in \mathcal{P}(\mathbb{R}^J)$ such that $L(\widehat{Q}) < \infty$. This is in applications often trivial to verify. A good candidate for $\widehat{Q}$ is typically the reference measure P.

Loss assumptions Let $\ell : \mathbb{R}^J \to \mathbb{R}$ be a loss satisfying the following assumptions:

(L1) The loss ℓ is bounded from below which means that

$$c := \inf\left\{\ell(\theta) : \theta \in \mathbb{R}^J\right\} > -\infty. \tag{C.2}$$

(L2) The loss is norm-coercive which means that

$$\ell(\theta) \to \infty \tag{C.3}$$

if $||\theta|| \to \infty$.

(L3) The loss ℓ is lower semi-continuous which means that

$$\liminf_{\theta \to \theta_0} \ell(\theta) \geq \ell(\theta_0) \tag{C.4}$$

for all $\theta_0 \in \mathbb{R}^J$.

Regulariser assumptions Let $D : \mathcal{P}(\mathbb{R}^J) \times \mathcal{P}(\mathbb{R}^J) \to [0, \infty]$ be a regulariser and $P \in \mathcal{P}(\mathbb{R}^J)$ a reference measure. We define $D_P(\cdot) := D(\cdot, P)$ for notational convenience. We assume the following for D_P:

(D1) The function D_P is lower semi-continuous w.r.t. to the topology of weak-convergence, i.e. for all sequences $\left(Q_n\right)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{R}^J)$ and all Q with $D_P(Q) < \infty$, it holds that $Q_n \xrightarrow{\mathcal{D}} Q$ implies

$$\liminf_{n\to\infty} D_P(Q_n) \geq D_P(Q). \tag{C.5}$$

Here, $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution.

(D2) D_P is strictly convex, i.e. for all $Q_1 \neq Q_2 \in \mathcal{P}(\mathbb{R}^J)$ with $D_P(Q_1) < \infty$ and $D_P(Q_2) < \infty$, it holds that

$$D_P\big(\alpha Q_1 + (1-\alpha)Q_2\big) < \alpha D_P(Q_1) + (1-\alpha)D_P(Q_2) \tag{C.6}$$

with $\alpha \in (0,1)$.

The next theorem provides an existence result for the optimisation problem (C.1). The result is similar in spirit to Lemma 2.1 in Knoblauch [2021] with the important difference that our assumptions are easier to verify, since they are formulated in terms of ℓ and D_P.

Theorem C.1.1 (Existence of global minimiser)**.** *Under the assumptions (L1)-(L3) and (D1) there exists a probability measure $Q^* \in \mathcal{P}(\mathbb{R}^J)$ with*

$$L(Q^*) = \inf\big\{L(Q) \,:\, Q \in \mathcal{P}(\mathbb{R}^J)\big\}. \tag{C.7}$$

Proof. Let $c > -\infty$ be the lower bound for ℓ. It follows immediately that $L(Q) \geq c$ for all $Q \in \mathcal{P}(\mathbb{R}^J)$ since $D(P,Q) \geq 0$. As a consequence we know that

$$\infty > L^* := \inf\big\{L(Q) \,:\, Q \in \mathcal{P}(\mathbb{R}^J)\big\} \geq c > -\infty. \tag{C.8}$$

By definition of the infimum we can construct a sequence $l_n = L(Q_n) \in \mathbb{R}$ in the image of L such

$$l_n \to L^* \tag{C.9}$$

for $n \to \infty$. We now show by contradiction that the corresponding sequence $\big(Q_n\big) \subset \mathcal{P}(\mathbb{R}^J)$ is *tight*[1]. Assume that $\big(Q_n\big)$ is not tight. By definition we can then find an $\epsilon > 0$ such that for each $k \in \mathbb{N}$ there exists $n = n_k \in \mathbb{N}$ with $Q_{n_k}([-k,k]^J) \leq 1 - \epsilon$. We set $A_k := [-k,k]^J \subset \mathbb{R}^J$ and obtain

$$\begin{aligned}
l_{n_k} &= L(Q_{n_k}) && \text{(C.10)}\\
&= \int_{A_k} \ell(\theta)\, dQ_{n_k}(\theta) + \int_{\mathbb{R}^J \setminus A_k} \ell(\theta)\, dQ_{n_k}(\theta) + \lambda D(Q,P) && \text{(C.11)}\\
&\geq \int_{A_k} \ell(\theta)\, dQ_{n_k}(\theta) + \int_{\mathbb{R}^J \setminus A_k} \ell(\theta)\, dQ_{n_k}(\theta) && \text{(C.12)}\\
&\geq c Q_{n_k}(A_k) + \inf\big\{\ell(\theta) \,:\, \theta \in \mathbb{R}^J \setminus A_k\big\} Q_{n_k}(\mathbb{R}^J \setminus A_k) && \text{(C.13)}
\end{aligned}$$

[1] A sequence of probability measures (Q_n) is called tight if and only if for every $\epsilon > 0$ there exists a compact set $K \in \mathbb{R}^J$ such that for all $n \in \mathbb{N}$ holds: $Q_n(K) > 1 - \epsilon$.

$$\geq cQ_{n_k}(A_k) + \epsilon \inf\left\{\ell(\theta) \,:\, \theta \in \mathbb{R}^J \backslash A_k\right\}. \tag{C.14}$$

Due to the coerciveness of ℓ, we know that $\inf\left\{\ell(\theta) \,:\, \theta \in \mathbb{R}^J \backslash A_k\right\} \to \infty$ for $k \to \infty$ and therefore $l_{n_k} \to \infty$ for $k \to \infty$. However, this is a contradiction: The sequence (l_n) is convergent and therefore in particular bounded. As a consequence, it cannot contain the unbounded sub-sequence (l_{n_k}). It follows that the sequence (Q_n) is tight. By Prokhorov's theorem we can now extract a sub sequence (Q_{n_k}) of (Q_n) and a measure $Q^* \in \mathcal{P}(\mathbb{R}^J)$ such that

$$Q_{n_k} \xrightarrow{\mathcal{D}} Q^* \tag{C.15}$$

for $k \to \infty$. Due to Lemma 5.1.7 in Ambrosio et al. [2005] the lower semi-continuity of ℓ implies that $Q \mapsto \int \ell(\theta)\, dQ(\theta)$ is lower semi-continuous. This combined with the lower semi-continuity of D_P gives

$$\liminf_{k\to\infty} L(Q_{n_k}) \geq L(Q^*). \tag{C.16}$$

From this it immediately follows that

$$L(Q^*) \leq \liminf_{k\to\infty} L(Q_{n_k}) = L^*, \tag{C.17}$$

but by definition L^* is the global minimum of L which implies $L^* \leq L(Q^*)$. We therefore conclude that $L(Q^*) = L^*$. □

Theorem C.1.1 only shows the existence of a global minimiser. In order to show uniqueness we use the convexity assumption (D2). The proof is the same as in finite dimensions and only included for completeness.

Theorem C.1.2 (Uniqueness of global minimiser)**.** *Assume that (D2) holds. Then, the global minimiser of L is unique (whenever it exists).*

Proof. Assume there exits two probability measures $Q_1, Q_2 \in \mathcal{P}(\mathbb{R}^J)$ such that

$$L(Q_1) = L^* = L(Q_2). \tag{C.18}$$

where $\infty > L^* := \inf\left\{L(Q) \,:\, Q \in \mathcal{P}(\mathbb{R}^J)\right\} > -\infty$. We define the probability measure $Q_3 := \frac{1}{2}Q_1 + \frac{1}{2}Q_3$. By strict convexity we obtain

$$L(Q_3) < \frac{1}{2}L(Q_1) + \frac{1}{2}L(Q_3) = L^*, \tag{C.19}$$

which is a contradiction to Q_1 and Q_2 being global minimisers. □

Note that in the literature on GVI [Knoblauch et al., 2019] it is common to assume that the regulariser is definite, i.e.

$$D(P, Q) = 0 \iff P = Q \tag{C.20}$$

for all $P, Q \in \mathcal{P}(\mathbb{R}^J)$. We did not use this assumption in neither Theorem C.1.1 nor Theorem C.1.2. However, the next lemma shows that it is basically implied by strict convexity.

Lemma C.1.3. *Let $D_P : \mathcal{P}(\mathbb{R}^J) \to [0, \infty]$ be strictly convex and assume further $D(Q, Q) = 0$ for all $Q \in \mathcal{P}(\mathbb{R}^J)$. Then it follows that $D(Q, P) = 0$ implies $P = Q$.*

Proof. We prove the claim by contradiction. Assume that there exists $P \neq Q$ such that $D(P, Q) = 0$. The strict convexity and $D(P, P) = 0$ imply combined that

$$D(\frac{1}{2}P + \frac{1}{2}Q, P) < \frac{1}{2}D(P, P) + \frac{1}{2}D(Q, P) \tag{C.21}$$

$$= 0. \tag{C.22}$$

However, we know that $D(\frac{1}{2}P + \frac{1}{2}Q, P) \geq 0$ by assumption. This is a contradiction. □

Discussion on loss assumptions The assumptions on the loss ℓ in (L1) and (L3) are rather weak. Typically loss functions in machine learning are bounded from below and continuous (and therefore in particular lower semi-continuous). However, norm-coercivity can be violated. Consider for example the squared loss

$$\ell(\theta) := \sum_{n=1}^{N} \big(y_n - f_\theta(x_n)\big)^2, \tag{C.23}$$

where f_θ is the parametrisation of a neural network with one hidden layer, i.e. $\theta = (w, A)$ and

$$f_\theta(x) = w^T \sigma(Ax), \tag{C.24}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function which is applied pointwise to the vector Ax and has the property that $\sigma(0) = 0$. It is now possible to find a sequence of parameters $(\theta_k)_{k\in\mathbb{N}} \subset \mathbb{R}^J$ with $||\theta_k|| \to \infty$ such that $\ell(\theta_k)$ does not converge to infinity. Define $w_k := k\big(1 \dots 1\big)$, $A_k := 0$ and $\theta_k = (w_k\ A_k)$ for $k \in \mathbb{N}$.

Then we obviously have that

$$||\theta_k|| = ||w_k|| \to \infty \tag{C.25}$$

for $k \to \infty$ but

$$\ell(\theta_k) = \sum_{n=1}^{N} \left(y_n - f_{\theta_k}(x_n)\right)^2 \tag{C.26}$$

$$= \sum_{n=1}^{N} \left(y_n - w^T \sigma(0)\right)^2 \tag{C.27}$$

$$= \sum_{n=1}^{N} y_n^2, \tag{C.28}$$

which is constant and therefore does not converge to ∞. A similar, but notationally more involved, construction can be made for neural networks with more than one hidden layer. However, this is an issue that can be easily resolved by adding what is known as weight decay to the loss. For example, consider for $\gamma > 0$ the loss

$$\ell(\theta) := \sum_{n=1}^{N} \left(y_n - f_\theta(x_n)\right)^2 + \gamma ||\theta||^2 \tag{C.29}$$

with weight decay. This loss is by construction norm-coercive and therefore the previous existence proof applies.

Discussion on regulariser assumptions The assumptions (D1) and (D2) are quite weak. The KL-divergence for example is known to be lower semi-continuous [Polyanskiy and Wu, 2014, Theorem 3.7] and strictly convex [Polyanskiy and Wu, 2014, Theorem 4.1]. This immediately implies lower semi-continuity and convexity of $\mathrm{KL}(\cdot, P)$ for any fixed P. The MMD is also known to be strictly convex [Arbel et al., 2019, Lemma 25], whenever it is well-defined, which can be guaranteed under weak assumptions on κ [Muandet et al., 2017, Lemma 3.1]. The lower semi-continuity properties also depend on the kernel κ. However, for bounded kernels it is trivial to verify. We include the proof for completeness, but assume this has been shown before elsewhere.

Lemma C.1.4. *Let the kernel $\kappa : \mathbb{R}^J \times \mathbb{R}^J$ be continuous and bounded: $||\kappa||_\infty := \sup_{\theta,\theta' \in \mathbb{R}^J} |k(\theta, \theta')| < \infty$ and P be fixed. Then $\mathrm{MMD}(\cdot, P)$ is continuous and therefore, in particular, lower semi-continuous.*

Proof. Let $(Q_n)_{n\in\mathbb{N}}$ and Q^* be such that

$$Q_n \xrightarrow{\mathcal{D}} Q^* \tag{C.30}$$

for $n \to \infty$. This immediately implies that

$$Q_n \otimes Q_n \xrightarrow{\mathcal{D}} Q^* \otimes Q^* \tag{C.31}$$

for $n \to \infty$, where $Q^* \otimes Q^*$ denotes the product measure of Q^* with itself. Further, note that the kernel mean embedding μ_P is continuous as integral with respect to the second component of a continuous function and bounded since

$$\begin{aligned} |\mu_P(\theta)| &= |\int \kappa(\theta, \theta')\, dP(\theta')| && \text{(C.32)} \\ &\leq \int |\kappa(\theta, \theta')| dP(\theta') && \text{(C.33)} \\ &\leq ||\kappa||_\infty. && \text{(C.34)} \end{aligned}$$

By the definition of weak convergence for measures, we therefore have

$$\begin{aligned} \iint \kappa(\theta, \theta')\, d(Q_n \otimes Q_n)(\theta, \theta') &\longrightarrow \iint \kappa(\theta, \theta')\, d(Q^* \otimes Q^*)(\theta, \theta') && \text{(C.35)} \\ \int \mu_P(\theta)\, dQ_n(\theta) &\longrightarrow \int \mu_P(\theta)\, dQ^*(\theta) && \text{(C.36)} \end{aligned}$$

for $n \to \infty$. This immediately implies continuity of $\mathrm{MMD}(\cdot, P)$ with respect to the topology of weak convergence. □

Notice that most kernels common in machine learning, such as the squared exponential or the Matérn kernel, are continuous and bounded and therefore Lemma C.1.4 applies.

Remark C.1.5. The astute reader may have noticed that our existence proof only guarantees the existence of measure $Q^* \in \mathcal{P}(\mathbb{R}^J)$. However, the Wasserstein gradient flow is by definition only formulated in the space of probability measures with finite second moment, denoted $\mathcal{P}_2(\mathbb{R}^J)$. Assumptions which guarantee that $Q^* \in \mathcal{P}_2(\mathbb{R}^J)$ are easy to formulate. For example, we can require that there exists $C > 0$ and $R > 0$ such that the loss ℓ satisfies

$$|\ell(\theta)| > C||\theta||^2 \tag{C.37}$$

for all $||\theta|| > R$. This immediately implies that $Q^* \in \mathcal{P}_2(\mathbb{R}^J)$ since otherwise

$$\int |\ell(\theta)|\, dQ^*(\theta) = \infty \tag{C.38}$$

gives a contradiction to the finiteness of $L(Q^*)$. However, even if (C.37) is violated, the reference measure

P may still guarantee that $Q^* \in \mathcal{P}_2(\mathbb{R}^J)$. For example, if $P \in \mathcal{P}_2(\mathbb{R}^J)$, then $D_P(Q^*)$ will typically be large if $Q^* \notin \mathcal{P}_2(\mathbb{R}^J)$ and the global minimiser is therefore in a sense *unlikely* to have fat tails. We therefore assume $Q^* \in \mathcal{P}_2(\mathbb{R}^J)$ throughout the paper and consider it to be a minor practical concern.

C.2 Realising the Wasserstein Gradient Flow

In this section, we identify a suitable stochastic process that allows us to follow the WGF.

Let $L^{\text{fe}} : \mathcal{P}(\mathbb{R}^J) \to (-\infty, \infty]$ be the free energy discussed in Section 5.3.2 given as

$$L^{\text{fe}}(Q) := \int V(\theta)\, dQ(\theta) + \frac{\lambda_1}{2} \int \kappa(\theta, \theta')\, dQ(\theta) dQ(\theta') + \lambda_2 \int \log\big(q(\theta)\big) q(\theta)\, d\theta, \tag{C.39}$$

where $\lambda_1, \lambda_2 \geq 0$ are constants, $V : \mathbb{R}^J \to \mathbb{R}$ is the potential, $\kappa : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}$ is symmetric. We will write L for L^{fe} from now on to simplify notation. The Wasserstein gradient of L is given as [cf. Chapter 9.1 Villani, 2003, Equation 9.4]

$$\nabla_W L[Q](\theta) = \nabla V(\theta) + \lambda_1 (\nabla_1 \kappa * Q)(\theta) + \lambda_2 \nabla \log\big(q(\theta)\big), \tag{C.40}$$

where $\nabla_1 \kappa : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}^J$ is the (vector-valued) derivative of κ with respect to the first component, ∇ denotes the euclidean gradient with respect to θ and $(\nabla_1 \kappa * Q)(\theta) := \int \nabla_1 \kappa(\theta, \theta')\, dQ(\theta')$ for $\theta \in \mathbb{R}^J$. The corresponding Wasserstein gradient flow is therefore given as [cf. Chapter 9.1 Villani, 2003, Equation 9.3]

$$\partial_t q(t, \theta) = \nabla \cdot \Big(q(t, \theta) \big(\nabla V(\theta) + \lambda_1 (\nabla_1 \kappa * Q)(\theta) + \lambda_2 \nabla \log\big(q_t(\theta)\big) \big) \Big). \tag{C.41}$$

In general the probability density evolution of a stochastic process is—via the Fokker-Planck equation—associated with the adjoint of the (infinitesimal) generator of the stochastic process. We will therefore try to identify the generator associated to the density evolution in (C.41). To this end let $h \in C_c^2(\mathbb{R}^J, \mathbb{R})$ where $C_c^2(\mathbb{R}^J, \mathbb{R})$ denotes the space of twice continuously differentiable functions with compact support. We multiply both sides of (C.41) with h, integrate, and apply the partial integration rule to obtain

$$\frac{d}{dt} \int h(\theta) q(t, \theta)\, d\theta = - \int \nabla_W L\big[Q(t)\big](\theta) \cdot \nabla h(\theta)\, q(t, \theta)\, d\theta. \tag{C.42}$$

$$= - \int \big(\nabla V(\theta) + \lambda_1 (\nabla_1 \kappa * Q_t)(\theta) \big) \cdot \nabla h(\theta)\, dQ_t(\theta) \tag{C.43}$$

$$- \lambda_2 \int \nabla \log\big(q_t(\theta)\big) \cdot \nabla h(\theta)\, dQ_t(\theta). \tag{C.44}$$

By chain-rule and partial integration, (C.44) can be rewritten as

$$-\lambda_2 \int \nabla \log\big(q_t(\theta)\big) \cdot \nabla h(\theta)\, dQ_t(\theta) = -\lambda_2 \int \nabla q_t(\theta) \cdot \nabla h(\theta)\, d\theta \tag{C.45}$$

$$= \lambda_2 \int \Delta h(\theta)\, dQ_t(\theta). \tag{C.46}$$

Putting everything together, we obtain

$$\frac{d}{dt} \int h(\theta) q(t,\theta)\, d\theta = \int \big(A[Q(t)]h\big)(\theta)\, dQ_t(\theta), \tag{C.47}$$

where $\big\{A[Q]\big\}_{Q \in \mathcal{P}(\mathbb{R}^J)}$ is a family of operators defined as

$$\big(A[Q]h\big)(\theta) := -\Big(\nabla V(\theta) + \lambda_1 (\nabla_1 \kappa * Q)(\theta)\Big) \cdot \nabla h(\theta) + \lambda_2 \Delta h. \tag{C.48}$$

for $h \in C_c^2(\mathbb{R}^J, \mathbb{R})$. The reader may recognize this operator family as the generator of a so called *nonlinear Markov processes* [Kolokoltsov, 2010, Chapter 1.4]. The nonlinearity in this case refers to the dependency on the measure Q. Linear Markov processes have no measure-dependency. This family of generators corresponds to a McKean-Vlasov process of the form

$$d\theta(t) = -\Big(\nabla V(\theta(t)) + \lambda_1 (\nabla_1 \kappa * Q_t)(\theta(t))\Big) dt + \sqrt{2\lambda_2} dB(t), \tag{C.49}$$

where $\big(B(t)\big)_{t>0}$ is a Brownian motion and Q_t the law of $\theta(t)$. In other words: The solution to (C.49) has the time marginals $Q(t)$ such that (C.47) holds for every $h \in C_c^2(\mathbb{R}^J, \mathbb{R})$. Furthermore, the corresponding pdfs $\big(q(t)\big)$ satisfy the nonlinear Fokker-Planck equation given as

$$\partial_t q_t = A^*[Q_t] q_t, \tag{C.50}$$

where $A^*[Q]$ denotes the L^2-adjoint of the operator $A[Q]$ and is given as

$$\big(A^*[Q]h\big)(\theta) = \nabla \cdot \Big(h(\theta)\big(\nabla V(\theta) + \lambda_1 (\nabla_1 \kappa * Q)(\theta) + \lambda_2 \nabla \log\big(h(\theta)\big)\big)\Big) \tag{C.51}$$

for $h \in C_c^2(\mathbb{R}^J, \mathbb{R})$[Barbu and Röckner, 2020, cf. equation (1.1)-(1.4)]. Note that (C.50) corresponds exactly to the Wasserstein gradient flow equation in (C.41). We can therefore follow the WGF by simulating solutions to (C.49).

The standard approach to simulate solutions to (C.49) [Veretennikov, 2006] is to use an ensemble of

interacting particles. Formally, we replace $Q(t)$ by $\frac{1}{N_E}\sum_{n=1}^{N_E}\delta_{\theta_n(t)}$ and obtain

$$d\theta_n(t) = -\Big(\nabla V\big(\theta_n(t)\big) + \frac{\lambda_1}{N_E}\sum_{j=1}^{N_E}(\nabla_1\kappa)\big(\theta_n(t),\theta_j(t)\big)\Big)dt + \sqrt{2\lambda_2}dB_n(t) \tag{C.52}$$

for $n = 1,\dots,N_E$ where $N_E \in \mathbb{N}$ denotes the number of particles. The Euler-Maruyama approximation of (C.52) leads to the final algorithm:

Step 1: Initialise $N_E \in \mathbb{N}$ particles $\theta_{1,0},\dots,\theta_{N_E,0}$ from a use chosen initial distribution Q_0.

Step 2: Evolve the particles forward in time according to

$$\theta_{n,k+1} = \theta_{n,k} - \eta\Big(\nabla V\big(\theta_{n,k}\big) + \frac{\lambda_1}{N_E}\sum_{j=1}^{N_E}(\nabla_1\kappa)\big(\theta_{n,k},\theta_{j,k}\big)\Big) + \sqrt{2\eta\lambda_2}Z_{n,k} \tag{C.53}$$

for $n = 1,\dots,N_E$, $k = 0,\dots,T-1$ with $Z_{n,k} \sim \mathcal{N}(0, I_{J\times J})$.

Note that $\theta_{n,k}$ is thought of as approximation of $\theta_n(t)$ at position $t = k\eta$. Furthermore, as discussed in Section 5.4, various choices of V, λ_1 and λ_2 allow us to implement the WGF for different regularised optimisation problems in the space of probability measures. This is summarised below:

- Deep ensembles: $V(\theta) = \ell(\theta)$, $\lambda_1 = 0$, $\lambda_2 = 0$
- Deep Langevin ensembles: $V(\theta) = \ell(\theta) - \lambda\log p(\theta)$, $\lambda_1 := 0$, $\lambda := \lambda_2$
- Deep repulsive Langevin ensembles: $V(\theta) = \ell(\theta) - \lambda_1\log p(\theta) - \lambda_2\mu_P(\theta)$

C.3 Asymptotic Distribution of Particles: Unregularised Objective

In this section, we investigate the asymptotic distribution of the WGF for the objective

$$L(Q) := \int \ell(\theta)\,dQ(\theta) \tag{C.54}$$

for $Q \in \mathcal{P}(\mathbb{R}^J)$. The associated particle method is:

- Sample $\theta_1(0),\dots,\theta_{N_E}(0)$ independently from Q_0.
- Simulate (deterministically) $\theta_n'(t) = -\nabla\ell\big(\theta_n(t)\big)$ for $n = 1,\dots,N_E$.

We start by introducing some notation for the deterministic gradient system. Let $\phi^t(\theta_0)$ denote the solution to the ordinary differential equation (ODE)

$$\theta(0) = \theta_0 \in \mathbb{R}^J \tag{C.55}$$

$$\theta'(t) = -\nabla \ell\big(\theta(t)\big) \tag{C.56}$$

at time $t > 0$. In a first step, we show the following lemma, which is a simple application of the famous Lojasiewicz theorem [Colding and Minicozzi II, 2014], and the fact that Lebesgue almost every initialisation leads to a local minimum [Lee et al., 2016].

Lemma C.3.1. *Assume $\ell : \mathbb{R}^J \to \mathbb{R}$ is norm-coercive and satisfies the Lojasiewicz inequality, i.e. for every $\theta \in \mathbb{R}^J$ exists an environment U of θ and constants $0 < \gamma < 1$ and $C > 0$ such that*

$$|\ell(\theta) - \ell(\bar{\theta})|^\gamma < C|\nabla \ell(\theta)|. \tag{C.57}$$

for all $\bar{\theta} \in U$. Then we know that $\phi^t(\theta_0)$ converges for $t \to \infty$ to a local minimum of ℓ for Lebesgue almost every $\theta_0 \in \mathbb{R}^J$.

Proof. First we show that $t \mapsto \phi^t(\theta_0)$ is bounded. We proof this by contradiction. Assume that $\phi^t(\theta_0)$ is unbounded. Then there exists a subsequence $(t_n)_{n\in\mathbb{N}} \subset [0, \infty)$ with $t_n \to \infty$ for $n \to \infty$ such that

$$|\phi^{t_n}(\theta_0)| \to \infty \tag{C.58}$$

for $n \to \infty$. The norm-coercivity immediately implies that

$$\ell\big(\phi^{t_n}(\theta_0)\big) \to \infty \tag{C.59}$$

for $n \to \infty$. However, this contradicts

$$\ell\big(\phi^t(\theta_0)\big) \leq \ell\big(\phi^0(\theta_0)\big) = \ell(\theta_0) < \infty, \tag{C.60}$$

where the first inequality follows from the fact that $t \mapsto \ell\big(\phi^t(\theta_0)\big)$ is decreasing, which is a consequence of

$$\frac{d}{dt}\ell\big(\phi^t(\theta_0)\big) = \nabla \ell\big(\phi^t(\theta_0)\big)\frac{d}{dt}\phi^t(\theta_0) \tag{C.61}$$

$$= -|\nabla \ell\big(\phi^t(\theta_0)\big)|^2 \leq 0. \tag{C.62}$$

Hence $t \mapsto \phi^t(\theta_0)$ is bounded. By the Bolzano-Weierstrass theorem we can find a sequence $(t_n)_{n\in\mathbb{N}} \subset [0,\infty)$ with $t_n \to \infty$ and a point $\theta_\infty \in \mathbb{R}^J$ such that

$$\phi^{t_n}(\theta_0) \to \theta_\infty \tag{C.63}$$

for $n \to \infty$. Hence $\left(\phi^t(\theta_0)\right)_{t>0}$ has the accumulation point θ_∞. The Lojasiewicz theorem [Colding and Minicozzi II, 2014] allows us to deduce that

$$\phi^t(\theta_0) \to \theta_\infty \tag{C.64}$$

for $t \to \infty$, and that θ_∞ satisfies $\nabla\ell(\theta_\infty) = 0$.

It remains to show that θ_∞ is not a saddle point for Lebesgue almost every initial value θ_0. However, this is very similar to the proof in Lee et al. [2016]. The only difference is that one would need to use a continuous-time version of the stable manifold theorem, which is readily available, for example in Bressan [2003]. □

Let $\{m_i\}_{i\in\mathbb{N}}$ denote the local minima of ℓ which are by assumption countable. Denote further by

$$\Theta_i := \left\{\theta_0 \in \mathbb{R}^J \,:\, \lim_{t\to\infty} \phi^t(\theta_0) \to m_i\right\} \tag{C.65}$$

the domain of attraction for the minimum m_i. The next theorem is then an easy consequence of Lemma C.3.1.

Theorem C.3.2. *Assume that the loss function ℓ only has countably many local minima, is norm coercive, and satisfies the Lojasiewicz inequality. Let further $\theta_0 \sim Q_0$ for some $Q_0 \in \mathcal{P}(\mathbb{R}^J)$ such that $\sum_{i=1}^\infty Q_0(\Theta_i) = 1$. Then,*

$$\phi^t(\theta_0) \xrightarrow{\mathcal{D}} \sum_{i=1}^\infty Q_0(\Theta_i)\,\delta_{m_i} =: Q_\infty \tag{C.66}$$

for $t \to \infty$. Here $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution.

Proof. Let $\theta_0 \in \mathbb{R}^J$ be fixed. Due to Lemma C.3.1, we know that

$$\phi^t(\theta_0) \to \sum_{i=1}^\infty m_i \mathbb{1}\{\theta_0 \in \Theta_i\} \tag{C.67}$$

for Lebesgue almost every θ_0 for $t \to \infty$. Here, $\mathbb{1}\{\cdot\}$ denotes the indicator function. Let Y now be a

random variable with law Q_0. By assumption, we know that $Y \in \Theta_i$ for some $i \in \mathbb{N}$ with probability 1. Hence,

$$\phi^t(Y) \to \sum_{i=1}^{\infty} m_i \mathbb{1}\{Y \in \Theta_i\} \tag{C.68}$$

almost surely for $t \to \infty$. Since almost sure convergence implies convergence in distribution, we conclude that

$$\phi^t(Y) \xrightarrow{\mathcal{D}} \mathcal{L}\big(\sum_{i=1}^{\infty} m_i \mathbb{1}\{Y \in \Theta_i\}\big), \tag{C.69}$$

where $\mathcal{L}(\cdot)$ denotes the law of a random variable. However, the law of the RHS is easily recognised as

$$\mathcal{L}\big(\sum_{i=1}^{\infty} m_i \mathbb{1}\{Y \in \Theta_i\}\big) = \sum_{i=1}^{\infty} Q_0(\Theta_i)\delta_{m_i}, \tag{C.70}$$

which concludes the proof. □

Remark C.3.3. Note that the condition

$$\sum_{i=1}^{\infty} Q_0(\Theta_i) = 1 \tag{C.71}$$

in Theorem C.3.2 is easy to satisfy. According to Lemma C.3.1 the set

$$\mathbb{R}^J \setminus \bigcup_{i=1}^{n} \Theta_i \tag{C.72}$$

has Lebesgue measure zero. Therefore, any Q_0 which has a density w.r.t. the Lebesgue measure will satisfy (C.71).

C.4 Asymptotic Distribution for Deep Langevin Ensembles

In this section, we analyse the objective

$$L(Q) := \int \ell(\theta)\, dQ(\theta) + \lambda\, \mathrm{KL}(Q, P) \tag{C.73}$$

for $Q \in \mathcal{P}(\mathbb{R}^J)$. The corresponding particle method is given as:

- Sample $\theta_1(0), \ldots, \theta_{N_E}(0)$ independently from Q_0.

- Simulate the SDE $d\theta_n(t) = -\nabla V\big(\theta_n(t)\big)dt + \sqrt{2\lambda}dB_n(t)$ for each $n = 1, \dots, N_E$.

Recall that $V(\theta) = \ell(\theta) - \lambda \log p(\theta)$. This case is well-studied in the literature and known as Langevin diffusion. Under mild assumptions [Chiang et al., 1987, Roberts and Tweedie, 1996],

$$\theta_n(t) \xrightarrow{\mathcal{D}} Q_\infty \tag{C.74}$$

for $t \to \infty$ and each particle $n = 1, \dots, N_E$ independently. The probability measure Q_∞ has the density

$$q_\infty(\theta) = \frac{1}{Z} \exp\big(-\frac{V(\theta)}{\lambda}\big) \tag{C.75}$$

$$= \frac{1}{Z} \exp\big(-\frac{\ell(\theta)}{\lambda}\big) p(\theta), \tag{C.76}$$

where $Z > 0$ is the normalising constant. As a consequence, the WGF asymptotically produces samples from Q_∞. However, it is a priori unclear that Q_∞ is in fact the same as the global minimiser Q^* of L.

We investigate this question by relating invariant measures to stationary points of the Wasserstein gradient.

Definition C.4.1. [Liggett, 2010, Thm. 3.3.7] A measure Q is called an invariant measure (for a given Feller-process) if

$$\int Ah(\theta)\, dQ(\theta) = 0 \tag{C.77}$$

for all $h \in C_c^2(\mathbb{R}^J)$. Here A is the infinitesimal generator of the corresponding Feller-process.

Recall that the infinitesimal generator of the Langevin diffusion for $h \in C_c^2(\mathbb{R}^J)$ is given as

$$Ah = -\nabla V \cdot \nabla h + \lambda \Delta h. \tag{C.78}$$

Definition C.4.2. A measure $Q \in \mathcal{P}_2(\mathbb{R}^J)$ is called a stationary point of the Wasserstein gradient if

$$\nabla_W L[Q](\theta) = 0 \tag{C.79}$$

for $Q-$almost every $\theta \in \mathbb{R}^J$.

In finite dimensions, it is well-known that a local minimiser is a stationary point of the gradient. This carries over to the infinite-dimensional case, with a similar proof. Since we could not find this result anywhere in the literature we included it for completeness.

Lemma C.4.3. *Let $\widehat{Q}$ be a local minimiser of L, i.e. there exits and $\epsilon > 0$ such that*

$$L(\widehat{Q}) \leq L(Q) \tag{C.80}$$

for all Q with $W_2(\widehat{Q}, Q) \leq \epsilon$. Then $\widehat{Q}$ is a stationary point of the Wasserstein gradient in the sense of Definition C.4.2.

Proof. Let $h \in C_c^2(\mathbb{R}^J)$ be arbitrary and $\widehat{Q} \in \mathcal{P}_2(\mathbb{R}^J)$ be a local minimum of L. Further, let $\phi^t(\theta_0)$ be the solution to the initial value problem

$$\theta(0) = \theta_0 \tag{C.81}$$

$$\theta'(t) = \nabla h(\theta(t)) \tag{C.82}$$

for $t \in (-\epsilon, \epsilon)$ for some $\epsilon > 0$. We now define $Q(t) := \phi^t \# \widehat{Q}$ for $t \in (-\epsilon, \epsilon)$ where $f \# \mu$ denotes the push-forward of the measures μ through the function f. In the Riemannian interpretation of the Wasserstein space, $\big(Q(t)\big)_{t \in (-\epsilon, \epsilon)}$ is a curve in $\mathcal{P}_2(\mathbb{R}^J)$ with tangent vector h at point $\widehat{Q}$ [Ambrosio et al., 2005, Chapter 8]. We, further, define $f : (-\epsilon, \epsilon) \to \mathbb{R}$ as $f(t) := L(Q(t))$. Application of the chain-rule [Ambrosio et al., 2005, p. 233] gives

$$f'(0) = \frac{d}{dt} L(Q(t))\big|_{t=0} \tag{C.83}$$

$$= \langle \nabla_W L[Q(0)], \nabla h \rangle_{L^2(Q(0))} \tag{C.84}$$

$$= \int \nabla_W L[\widehat{Q}](\theta) \cdot \nabla h(\theta) \, d\widehat{Q}(\theta). \tag{C.85}$$

We know that f has a local minimum at $t = 0$ and, therefore, $f'(0) = 0$ which gives

$$0 = \int \nabla_W L[\widehat{Q}](\theta) \cdot \nabla h(\theta) \, d\widehat{Q}(\theta). \tag{C.86}$$

Since (C.86) holds for arbitrary test functions $h \in C_c^2(\mathbb{R}^J)$ and as $C_c^2(\mathbb{R}^J)$ is dense in $L^2(\widehat{Q})$, we obtain that $\nabla_W L[\widehat{Q}](\theta) = 0$ for $\widehat{Q}$-a.e $\theta \in \mathbb{R}^J$. □

The next lemma relates invariant measures and stationary points of the Wasserstein gradient for infinitesimal generators of the form (C.78). It will prove extremely useful to translate between the Langevin diffusion literature and our optimisation perspective.

Lemma C.4.4. *Let $Q \in \mathcal{P}_2(\mathbb{R}^J)$ be such that Q has a density q with respect to the Lebesgue measure. Then, the following two statements are equivalent:*

- *Q is a stationary point of the Wasserstein gradient.*
- *Q is an invariant measure.*

Proof. Let Q be a measure with density q. Recall that the generator of the Langevin diffusion is for $h \in C_c^2(\mathbb{R}^J)$ given as

$$Ah = -\nabla V \cdot \nabla h + \lambda \Delta h. \tag{C.87}$$

By partial integration, it is easy to verify that the L^2- adjoint (w.r.t the Lebesgue measure) is given as

$$A^* h = \nabla \cdot (h \cdot \nabla V) + \lambda \Delta h. \tag{C.88}$$

We, therefore, conclude that

$$\begin{aligned} \int Ah(\theta)\, dQ(\theta) &= \int Ah(\theta) q(\theta)\, d\theta && \text{(C.89)} \\ &= \int h(\theta) A^* q(\theta)\, d\theta && \text{(C.90)} \\ &= \int h(\theta) \Big(\nabla \cdot (q(\theta) \cdot \nabla V(\theta)) + \lambda \Delta q(\theta) \Big)\, d\theta. && \text{(C.91)} \end{aligned}$$

Furthermore, we have $\nabla_W L[Q] = \nabla V + \lambda \nabla \log q$, and therefore

$$\begin{aligned} \int \nabla_W L[Q](\theta) \cdot \nabla h(\theta)\, dQ(\theta) &= \int \nabla_W L[Q](\theta) \cdot \nabla h(\theta) q(\theta)\, d\theta && \text{(C.92)} \\ &= \int \Big(\nabla V(\theta) q(\theta) + \lambda \nabla q(\theta) \Big) \cdot \nabla h(\theta)\, d\theta && \text{(C.93)} \\ &= - \int h(\theta) \Big(\nabla \cdot (q(\theta) \nabla V(\theta)) + \lambda \Delta q(\theta) \Big)\, d\theta, && \text{(C.94)} \end{aligned}$$

where the last line follows from applying partial integration. This allows us to conclude that

$$\int Ah(\theta)\, dQ(\theta) = - \int \nabla_W L[Q](\theta) \cdot \nabla h(\theta)\, dQ(\theta) \tag{C.95}$$

whenever Q has a density. As a consequence we have that Q is invariant if and only if it is a stationary point of the Wasserstein gradient. □

Lemma C.4.4 allows us to move between the optimisation and stochastic differential equation perspective. In Appendix C.1, we discussed the existence and uniqueness of a global minimiser Q^* of L. We know that Q^* has a density since the Kullback-Leibler divergence would be infinite otherwise (assuming P

has a Lebesgue-density which we assume throughout the paper). Lemma C.4.3 guarantees that Q^* is a stationary point of the Wasserstein gradient. Due to Lemma C.4.4, we can infer that Q^* must be an invariant measure. However, due to the uniqueness of the invariant measure under the previously mentioned mild assumptions [Chiang et al., 1987, Roberts and Tweedie, 1996], we can conclude that $Q^* = Q_\infty$.

C.5 Asymptotic Distribution of Deep Repulsive Langevin Ensembles

This section contains the proof of Theorem 5.4.2 which is repeated hereafter.

Theorem C.5.1. *Let $Q^{n,N_E}(t)$ be the distribution of $\theta_n(t)$, $n = 1, \dots, N_E$, generated via (5.7). Then*

$$\lim_{t\to\infty} \lim_{N_E\to\infty} Q^{n,N_E}(t) = Q^* \text{ (in distribution)}$$

for each $n = 1, \dots, N_E$ whenever the corresponding the McKean-Vlasov process converges to a unique invariant measure.

Proof. The ergodicity of McKean-Vlasov processes has been studied by several authors [Veretennikov, 2006, Mishura and Veretennikov, 2020, Liu and Ma, 2022]. The conditions for ergodicity involve a variety of requirements on the drift and diffusion terms, which are too numerous to list here. These conditions translate into constraints on the kernel κ and the potential $V(\theta) = \ell(\theta) - \lambda_1 \log p(\theta) + \lambda_2 \mu_P(\theta)$. In our case, the McKean-Vlasov process describing the Wasserstein gradient flow is given by

$$\theta(0) \sim Q_0 \tag{C.96}$$

$$d\theta(t) = -\Big(\nabla V(\theta(t)) + \lambda_1 (\nabla_1 \kappa * Q_t)(\theta(t))\Big) dt + \sqrt{2\lambda_2} dB(t), \tag{C.97}$$

with $\big(B(t)\big)_{t\geq 0}$ being a Brownian motion realises the Wasserstein gradient flow. For example, Section 2 of Veretennikov [2006] provides an extensive list of conditions that closely align with our situation. Under these assumptions, we know from Theorem 2 in Veretennikov [2006] that

$$Q^{n,N_E}(t) \xrightarrow{\mathcal{D}} Q_\infty \quad \text{(first as } N_E \to \infty \text{, then as } t \to \infty), \tag{C.98}$$

where Q_∞ is the invariant measure of the McKean-Vlasov SDE. The remaining task is to show that Q_∞ coincides with the global minimizer Q^*.

We start by introducing the concept of an invariant measure for a nonlinear Markov process [Ahmed and Ding, 1993, Definition 1].

Definition C.5.2. A measure Q is called an invariant measure for a nonlinear Markov process with the family of infinitesimal generators $\{A[Q]\}_{Q\in\mathcal{P}(\mathbb{R}^J)}$ if

$$\int A[Q]h(\theta)\,dQ(\theta) = 0 \tag{C.99}$$

for all $h \in C_c^2(\mathbb{R}^J)$.

Recall that L in this section is given as

$$L(Q) := \int V(\theta)\,dQ(\theta) + \frac{\lambda_1}{2}\iint \kappa(\theta,\theta')\,dQ(\theta)dQ(\theta') + \lambda_2 \int \log q(\theta)q(\theta)\,d\theta, \tag{C.100}$$

and the infinitesimal generators of the McKean-Vlaslov SDE is given as

$$\big(A[Q]h\big)(\theta) := -\Big(\nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta)\Big)\cdot\nabla h(\theta) + \lambda_2\Delta h. \tag{C.101}$$

for $h \in C_c^2(\mathbb{R}^J,\mathbb{R})$. In analogy to Lemma C.4.4, we obtain the following result.

Lemma C.5.3. *Let $Q \in \mathcal{P}_2(\mathbb{R}^J)$ be such that Q has a density q with respect to the Lebesgue measure. Then, the following two statements are equivalent:*

- *Q is a stationary point of the Wasserstein gradient for L in (C.100) in the sense of Def. C.4.2.*
- *Q is an invariant measure for the McKean-Vlasov process with infinitesimal generator defined in (C.101)*

First, we notice that

$$\int A[Q]h(\theta)dQ(\theta) = \int A[Q]h(\theta)q(\theta)\,d\theta \tag{C.102}$$

$$= \int h(\theta)\big(A^*[Q]q\big)(\theta)\,d\theta. \tag{C.103}$$

Recall, that $A^*[Q]$ denotes the L^2-adjoint of the operator $A[Q]$ and that it is given as

$$\big(A^*[Q]h\big)(\theta) = \nabla\cdot\Big(h(\theta)\big(\nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta) + \lambda_2\nabla\log\big(h(\theta)\big)\big)\Big) \tag{C.104}$$

for $h \in C^2(\mathbb{R}^J,\mathbb{R})$ with compact support. This implies

$$\big(A^*[Q]q\big)(\theta) = \nabla\cdot\Big(q(\theta)\big(\nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta)\big)\Big) + \lambda_2\Delta q(\theta). \tag{C.105}$$

We plug this into (C.103) to obtain

$$\int A[Q]h(\theta)\,dQ(\theta) \tag{C.106}$$

$$= \int h(\theta)\,\nabla\cdot\Big(q(\theta)\big(\nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta)\big)\Big)\,d\theta + \int \lambda_2 h(\theta)\Delta q(\theta)\,d\theta. \tag{C.107}$$

On the other hand, we have that

$$\nabla_W L[Q](\theta) = \nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta) + \lambda_2\nabla\log q(\theta), \tag{C.108}$$

and therefore

$$\int \nabla L[Q](\theta)\cdot\nabla h(\theta)\,dQ(\theta) \tag{C.109}$$

$$= \int \Big(\nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta) + \lambda_2\nabla\log q(\theta)\Big)\cdot\nabla h(\theta)\,dQ(\theta) \tag{C.110}$$

$$= \int q(\theta)\big(\nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta)\big)\cdot\nabla h(\theta)\,d\theta + \lambda_2\int \nabla q(\theta)\cdot\nabla h(\theta)\,d\theta \tag{C.111}$$

$$= -\int \nabla\cdot\Big(q(\theta)\big(\nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta)\big)\Big)h(\theta)\,d\theta - \lambda_2\int q(\theta)\Delta h(\theta)\,d\theta, \tag{C.112}$$

where the last line follows from partial integration. Comparing (C.107) to (C.112) gives

$$\int A[Q]h(\theta)\,dQ(\theta) = -\int \nabla L[Q](\theta)\cdot\nabla h(\theta)\,dQ(\theta) \tag{C.113}$$

for all $h \in C_c^2(\mathbb{R}^J)$ whenever Q has a density. This immediately implies that Q is invariant iff it is a stationary point.

Let now Q^* be the unique global minimiser of L. Due to Lemma C.4.3, we know that Q^* is a stationary point of the Wasserstein gradient. Lemma C.5.3 shows that Q^* is the invariant measure of the McKean-Vlaslov SDE which is what we had to show.

□

C.6 Asymptotic Analysis of Deep Repulsive Ensembles

In this section, we consider the objective

$$L(Q) := \int \ell(\theta)\,dQ(\theta) + \lambda\,\mathrm{MMD}(Q, P) \tag{C.114}$$

for $Q \in \mathcal{P}(\mathbb{R}^J)$. The corresponding McKean-Vlasov process is of the form

$$d\theta(t) = -\Big(\nabla V(\theta(t)) + \lambda(\nabla_1 \kappa * Q_t)(\theta(t))\Big)dt, \tag{C.115}$$

where Q_t denotes the distribution of $\theta(t)$ and $V(\theta) = \ell(\theta) - \mu_P(\theta)$ with $\mu_P(\theta) = \int \kappa(\theta, \theta')\, dP(\theta)$ the kernel mean-embedding of P. We call the particle method in this case **deep repulsive ensembles (DRE)**.

The existence of the global minimiser Q^* is still guaranteed under the assumptions in Appendix C.1. Lemma C.4.3 guarantees that Q^* is a stationary point of the Wasserstein gradient, i.e.

$$\nabla V(\theta) + \lambda(\nabla_1 \kappa * Q^*)(\theta) = 0 \tag{C.116}$$

for Q^*-a.e. $\theta \in \mathbb{R}^J$. Recall that the infinitesimal generator in this case is given as

$$\big(A[Q]h\big)(\theta) := -\Big(\nabla V(\theta) + \lambda(\nabla_1 \kappa * Q)(\theta)\Big) \cdot \nabla h(\theta) \tag{C.117}$$

for $Q \in \mathcal{P}(\mathbb{R}^J)$, $h \in C_c^2(\mathbb{R}^J)$. It immediately follows from the definition that

$$\big(A[Q]h\big)(\theta) = -\nabla_W L[Q](\theta) \cdot \nabla h(\theta) \tag{C.118}$$

for all $h \in C_c^2(\mathbb{R}^J)$, $\theta \in \mathbb{R}^J$. As in Lemma C.4.4 & C.5.3, this implies that each stationary point of the Wasserstein gradient is an invariant measure of the McKean-Vlasov process and vice versa. In Appendix C.4 & C.5, we cite relevant literature that guarantees uniqueness of the invariant measure, which is a necessary (but not sufficient) condition for convergence to the invariant measure. The next theorem shows that uniqueness will in general not hold without the presence of the diffusion term.

Theorem C.6.1. *The invariant measure for the McKean-Vlasov process with the family of generators* $\big(A[Q]\big)_{Q \in \mathcal{P}(\mathbb{R}^J)}$ *defined in* (C.118) *is (in general) not unique.*

Proof. Let $N_E \in \mathbb{N}$ and define $\widetilde{L} : \big(\mathbb{R}^J\big)^{N_E} \to \mathbb{R}$ as

$$\widetilde{L}(\theta_1, \dots, \theta_{N_E}) := \sum_{i=1}^{N_E} V(\theta_i) + \frac{\lambda}{2N_E} \sum_{i,j=1}^{N_E} \kappa(\theta_i, \theta_j). \tag{C.119}$$

Assume that V is bounded from below and norm-coercive. Then $\widetilde{L}$ is bounded from below and norm-coercive and therefore we can find a global minimiser $\theta^* := \big(\theta_1^*, \dots, \theta_{N_E}^*\big) \in \big(\mathbb{R}^J\big)^{N_E}$ of $\widetilde{L}$. Since $\widetilde{L}$ is

differentiable, we know that θ^* is a stationary point of the gradient which implies

$$\nabla V(\theta_i^*) + \frac{\lambda}{N_E} \sum_{j=1}^{N_E} (\nabla_1 \kappa)(\theta_i^*, \theta_j^*) = 0 \tag{C.120}$$

for all $i = 1, \dots, N_E$. Here, we assume that the kernel κ is symmetric, which is standard in the MMD literature. Note that (C.120) is equivalent to

$$\nabla V(\theta) + \lambda (\nabla_1 \kappa * \widehat{Q})(\theta) = 0 \tag{C.121}$$

for $\widehat{Q}$-a.e. $\theta \in \mathbb{R}^J$ where

$$\widehat{Q}(d\theta) := \frac{1}{N_E} \sum_{j=1}^{N_E} \delta_{\theta_j^*}(d\theta). \tag{C.122}$$

This means that $\widehat{Q}$ is a stationary point of the Wasserstein gradient, and therefore an invariant measure for the McKean-Vlasov process. Since $N_E \in \mathbb{N}$ was arbitrary, we have constructed countably many invariant measures and therefore uniqueness can't hold in general. □

The reason that non-uniqueness of the invariant measure is an immediate contradiction to convergence is the following: If we initialise with any of the invariant measures constructed in the proof of Theorem C.6.1, then the particle distribution of the McKean-Vlasov process will remain unchanged over time. Convergence to the global minimiser can therefore surely not hold for arbitrary initialisation Q_0. It may be possible to construct conditions on Q_0 under which convergence still holds. For example, for Stein variational gradient descent a similar issue occurs. However, in this case one can guarantee convergence [Lu et al., 2019, Theorem 2.8] if Q_0 has a Lebesgue-density (and if the kernel satisfies further restrictive assumptions). The existence of conditions that guarantee convergence for DRE remains an open problem.

C.7 Implementation Details

In Appendix C.1, we derived the following algorithm:

Step 1: Simulate $N_E \in \mathbb{N}$ particles $\theta_{1,0}, \dots, \theta_{N_E,0}$ from a use chosen initial distribution Q_0.

Step 2: Evolve the particles forward in time according to

$$\theta_{n,k+1} = \theta_{n,k} - \eta\Big(\nabla V\big(\theta_{n,k}\big) + \frac{\lambda_1}{N_E} \sum_{j=1}^{N_E} (\nabla_1 \kappa)\big(\theta_{n,k}, \theta_{j,k}\big)\Big) + \sqrt{2\eta\lambda_2} Z_{n,k} \tag{C.123}$$

for $n = 1, \dots, N_E$, $k = 0, \dots, K-1$ with $Z_{n,k} \sim \mathcal{N}(0, I_{J\times J})$.

We can generate samples from DE, DLE and DRLE by setting the potential and regularisation parameters as described below:

- Deep ensembles: $V(\theta) = \ell(\theta)$, $\lambda_1 = 0$, $\lambda_2 = 0$
- Deep Langevin ensembles: $V(\theta) = \ell(\theta) - \lambda \log p(\theta)$, $\lambda_1 = 0$, $\lambda = \lambda_2$
- Deep repulsive Langevin ensembles: $V(\theta) = \ell(\theta) - \lambda_1 \log p(\theta) - \lambda_2 \mu_P(\theta)$

Due to Appendix C.4 & C.5, we can think of $\theta_{1,K}, \dots, \theta_{N_E,K}$ as approximately sampled from the global minimiser Q^* for DLE and DRLE if K is large enough. All experiments use the SE kernel given as

$$\kappa(\theta, \theta') = \exp\big(-\frac{||\theta - \theta'||^2}{2\sigma_\kappa^2}\big) \tag{C.124}$$

with lengthscale parameter $\sigma_\kappa > 0$. The kernel mean embedding μ_P can easily be approximated as

$$\mu_P(\theta) = \frac{1}{M}\sum_{i=1}^{M} \kappa(\theta, \theta_i),\ \theta \in \mathbb{R}^J, \tag{C.125}$$

where $\theta_1, \dots, \theta_M \sim P$ independently. We chose $M = 20$.

C.7.1 Toy Example: Global Minimiser

We describe details regarding the experiments conducted to produce Figure 5.2 below.

We generate $N_E = 300$ particles and make the following choices:

- Loss: $\ell(\theta) := \frac{3}{2}(\frac{1}{4}\theta^4 + \frac{1}{3}\theta^3 - \theta^2) - \frac{3}{8}$
- Prior: $P \sim \mathcal{N}(0, 1)$ and therefore $\log p(\theta) = -\frac{1}{2}\theta^2$
- Initialisation: $Q_0 = P$
- Reg. parameter: $\lambda_{DLE} = 1$, $\lambda_{DRLE} = 1$, $\lambda'_{DRLE} = 1$
- Step size: $\eta = 10^{-4}$, Iterations: $K = 100,000$
- Kernel lengthscale, σ_κ, is chosen according to the median heuristic [Garreau et al., 2017] based on samples from the prior P

The loss is constructed such that we have a global minimum at $\theta = -2$, a turning point at $\theta = 0$, and a local minimum at $\theta = 1$.

- Deep ensembles: The optimal Q^* is a Dirac measure located at the global minimiser $\theta = -2$. However, as we proved in Theorem 5.4.1, the WGF produce samples from

$$Q_\infty(d\theta) = \frac{1}{2}\delta_{-2}(d\theta) + \frac{1}{2}\delta_1(d\theta), \tag{C.126}$$

as $(-\infty, 0)$ is the region of attraction for the global minimum and $(0, \infty)$ for the local minimum which both have probability 0.5 under $Q_0 = P = \mathcal{N}(0, 1)$. In particular, $Q_\infty \neq Q^*$ as expected.

- Deep Langevin ensembles: The optimal measure has the pdf

$$q^*(\theta) \propto \exp\big(-\frac{\ell(\theta)}{\lambda}\big)p(\theta) \tag{C.127}$$

for $\theta \in \mathbb{R}$. As expected the WGF produces samples from Q^*.

- Deep repulsive Langevin ensembles: The optimal q^* for deep repulsive ensembles is harder to determine. From the condition that q^* is a stationary point of the Wasserstein gradient, we can derive that $u(\theta) := \log q^*(\theta)$ satisfies the integro-differential equation

$$u'(\theta) = -\frac{1}{\lambda_2}V'(\theta) - \frac{\lambda_1}{\lambda_2}\int (\nabla_1\kappa)(\theta, \theta')\exp(u(\theta'))\, d\theta' \tag{C.128}$$

with some initial value $u(0) = u_0$. In principle, we could choose u_0 such that $q(\theta) := \exp(u(\theta))$ integrates to 1. However, since we do not know the appropriate initial condition a priori, we choose an arbitrary u_0 and normalise the pdf afterwards. We use an numerical solver to evaluate $u(\theta)$ on a fixed grid. As expected, the WGF produces samples from Q^* in this case.

C.7.2 Toy Example: Multimodal Loss

The details below correspond to the experimental results presented in Figures 5.3 and C.1. Figure C.1 is an alteration of Figure 5.3 with only 4 particles with the goal of making the messaging of the number of particles relative to the local minima clearer.

DE, DLE, DRLE We generate $N_E = 300$ particles and make the following choices:

- Loss: $\ell(\theta) = -\log\sum_{i=1}^4 \frac{1}{4}\mathcal{N}(\theta; \mu_i, I_2)$, $\theta \in \mathbb{R}^2$, $\mu_i = (\pm 3, \pm 3)^T$, $i = 1, \dots, 4$
- Prior: P flat and therefore $\log p(\theta) = 0$
- Initialisation: $Q_0 \sim \mathcal{N}(0, I_2)$
- Reg. parameter: $\lambda_{DLE} = 0.2$, $\lambda_{DRLE} = 0.2$, $\lambda'_{DRLE} = 0.6$

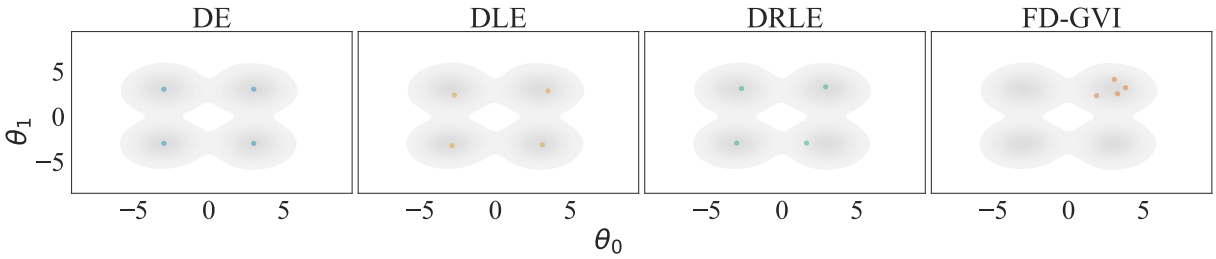


Figure C.1: We generate $N_E = 4$ particles from DE, DLE, DRLE and FD-GVI with Gaussian parametrisation. The multimodal loss ℓ is plotted in grey and the particles of the different methods are layered on top. The prior in this example is flat, i.e. $\log p$ and μ_P are constant. The initialisation Q_0 is standard Gaussian.

- Step size: $\eta = 0.1$, Iterations: $K = 10,000$
- Kernel lengthscale, σ_κ, is chosen according to the median heuristic [Garreau et al., 2017] based on samples from the prior P

Note that for a translation-invariant kernel such as the SE kernel we obtain for the flat prior P that

$$\mu_P(\theta) = \int_{-\infty}^{\infty} \kappa(\theta, \theta')\, d\theta' \tag{C.129}$$

$$= \int_{-\infty}^{\infty} \phi(\theta - \theta')\, d\theta' \tag{C.130}$$

$$= \int_{-\infty}^{\infty} \phi(\xi)\, d\xi, \tag{C.131}$$

where the second line follows from the fact that we can write any translation-invariant kernel as $\kappa(\theta, \theta') = \phi(\theta - \theta')$ for some function $\phi : \mathbb{R}^J \to \mathbb{R}$ and the second line is simple variable substitution. If (C.131) is finite, the above expression is well-defined and therefore μ_P constant. Note that in particular for the SE kernel, we have $\phi(\xi) = \exp(-||\xi||^2/(2\sigma_\kappa^2))$ and therefore (C.131) is finite. As a consequence, we have that for a flat prior P the gradient of the potential V is the same for all three methods. This means that the loss ℓ isn't adjusted and the only difference between the three methods is the presence of repulsion and noise effects.

Remark C.7.1. The astute reader may have noticed that a flat prior P is in fact not covered by our theory in Appendix C.1. The problem is that $\mathrm{KL}(\cdot, \mathcal{L})$, where $\mathcal{L}$ denotes the Lebesgue measure, is not positive (and not even bounded from below). To see this, choose $Q = \mathcal{N}(0, \Sigma)$ with $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$ and note that

$$\mathrm{KL}(Q, \mathcal{L}) = \int \log q(\theta)\, q(\theta)\, d\theta = -\mathrm{H}(Q), \tag{C.132}$$

where $\mathrm{H}(Q)$ denotes the differential entropy. For a Gaussian, it is known that

$$\mathrm{H}(Q) = \frac{1}{2}\log\left((2\pi e)^J \det(\Sigma)\right) = \frac{1}{2}\left(\log(2\pi e)^J + \log(\sigma_1^2) + \log(\sigma_2^2)\right) \tag{C.133}$$

and therefore if either $\sigma_1^2 \to \infty$ or $\sigma_2^2 \to \infty$ then $\mathrm{KL}(Q, \mathcal{L}) \to -\infty$. However, note that this difficulty is rather technical in nature and can easily be remedied. Instead of $\mathcal{L}$, we could have chosen the uniform prior $P \sim U(-10^{100}, 10^{100})$. In this case, the positivity of $\mathrm{KL}(\cdot, P)$ is guaranteed by Jensen's inequality. This choice of P gives—up to an additive constant—the same objective as a flat prior and up to machine precision the same kernel mean embedding μ_P. It is, therefore, algorithmically irrelevant if P is flat or uniform on a very large set.

FD-GVI We use the same prior and loss as for DE, DLE and DRLE. We parameterise the variational family as independent Gaussian, i.e.

$$\mathcal{Q} = \left\{\mathcal{N}(\mu, \Sigma) \,|\, \mu \in \mathbb{R}^2,\, \Sigma = \mathrm{diag}\left(\exp(\beta_1), \exp(\beta_2)\right), \beta := (\beta_1, \beta_2)^2 \in \mathbb{R}^2\right\}. \tag{C.134}$$

We learn the variational parameters $\nu := (\mu, \beta) \in \mathbb{R}^4$ by minimising

$$\widetilde{L}(\nu) = \int \ell(\theta)\, dQ_\nu(\theta) + \lambda KL(Q_\nu, P) \tag{C.135}$$

$$= \int \ell(\theta)\, dQ_\nu(\theta) - \lambda H\left(\mathcal{N}(\mu, \Sigma)\right) \tag{C.136}$$

$$\approx \frac{1}{200}\sum_{j=1}^{200} \ell(\mu + \Sigma^{0.5} Z_j) - \frac{\lambda}{2}\log\left((2\pi e)^2 \exp(\beta_1)\exp(\beta_2)\right) \tag{C.137}$$

$$= \frac{1}{200}\sum_{j=1}^{200} \ell(\mu + \Sigma^{0.5} Z_j) - \frac{\lambda}{2}\left(\beta_1 + \beta_2\right) + const, \tag{C.138}$$

where $Z_1, \ldots, Z_{200} \sim \mathcal{N}(0, I_2)$ and $H\left(\mathcal{N}(\mu, \Sigma)\right)$ denotes the differential entropy of the normal distribution. For the regularisation parameter, we chose $\lambda = 0.5$.

C.7.3 Toy Example: More Modes than Particles

We generate $N_E = 20$ particles and make the following choices:

- Loss: $\ell(\theta) = -|\sin(\theta)|$, $\theta \in [-M\pi, M\pi]$, with $M = 1000$
- Prior: P flat and therefore $\log p(\theta) = 0$ and $\mu_P = \text{const.}$ (cf. Appendix C.7.2)
- Initialisation: $Q_0 \sim U(-M\pi, M\pi)$
- Reg. parameter: $\lambda_{DLE} = 0.001$, $\lambda_{DRLE} = 0.001$, $\lambda'_{DRLE} = 0.6$

- Step size: $\eta = 0.01$, Iterations: $K = 1.000$
- Kernel lengthscale, σ_κ, is chosen according to the median heuristic [Garreau et al., 2017] based on samples from the prior P

Note that ℓ has $2M = 2000$ local minima at locations

$$m_i := \frac{\pi}{2} + i\pi, \quad i \in \{-M, \ldots, 0, \ldots, (M-1)\}. \tag{C.139}$$

Due to the flat prior $\nabla V = \nabla \ell$ for all three methods. We observe that it is hard to distinguish the methods since most particles are in their local modes by themselves.

C.7.4 UCI Regression

The UCI data sets are licensed under Creative Commons Attribution 4.0 International license (CC BY 4.0). Following Lakshminarayanan et al. [2017], we train 5 one-hidden-layer neural networks f_θ with 50 hidden nodes for 40 epochs. We split each data set into train (81% of samples), validation (9% of samples), and test set (10% of samples). Based on the best hyperparameter runs (according to a Gaussian NLL) found via grid search on a validation data set, we make the following choices:

- Loss: $\ell(\theta) = \frac{1}{N}\sum_{n=1}^{N}(f_\theta(x_n) - y_n)^2$ where $\{x_n, y_n\}_{n=1}^{N}$ are paired observations.
- Prior: $P \sim \mathcal{N}(0, 1)$
- Initialisation: Kaiming intilisation, i.e. for each layer $l \in \{1, ...L\}$ that maps features with dimensionality n_{l-1} into dimensionality n_l, we sample $Q_{l,0} \sim \mathcal{N}(0, 2/n_l)$
- Reg. parameter: $\lambda_{DLE} = 10^{-4}$, $\lambda_{DRLE} = 10^{-4}$, $\lambda'_{DRLE} = 10^{-2}$
- Step size: $\eta = 0.1$, Iterations: $K = 10,000$
- Kernel lengthscale, σ_κ, is chosen according to the median heuristic [Garreau et al., 2017] based on samples from the prior P

C.7.5 Compute

While the final experimental results can be run within approximately an hour on a single GeForce RTX 3090 GPU, the complete compute needed for the final results, debugging runs, and sweeps amounts to around 9 days.

D | Bayesian Inference in Function Space via the Wasserstein Gradient Flow

D.1 Technical Background

D.1.1 Reproducing Kernel Hilbert Spaces

Let $\mathcal{X}$ be non-empty set and $\mathcal{F}(\mathcal{X}, \mathbb{R})$ be the set containing all functions $f : \mathcal{X} \to \mathbb{R}$.

A Hilbert space $\big(H, \langle \cdot, \cdot \rangle\big)$ with $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ is called Reproducing Kernel Hilbert Space (RKHS) if and only if the pointwise evaluation functionals $\pi_x : H \to \mathbb{R}$, $f \mapsto f(x)$ are continuous for all $x \in \mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called kernel, if the kernel matrix $k(x_{1:N}, x_{1:N}) := \big(k(x_n, x_{n'})\big)_{n,n'=1}^N \in \mathbb{R}^{N \times N}$ is positive semi-definite for all $x_{1:N} \in \mathcal{X}^N$ and $N \in \mathbb{N}$.

Moore-Aronszajn theorem [Aronszajn, 1950] states that, for every RKHS H, there exists a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the property that $k(x, \cdot) \in H$ and $\pi_x(f) = \big\langle f, k(x, \cdot) \big\rangle$ for all $x \in \mathcal{X}$. The functions $k(x, \cdot) : \mathcal{X} \to \mathbb{R}$ are called canonical feature maps and the property $f(x) = \langle f, k(x, \cdot) \rangle$ is called reproducing property.

Conversely, for any kernel k we can construct a Hilbert space $H \subset \mathcal{F}(\mathcal{X}, \mathbb{R})$ as closure of H_0 defined as

$$H_0 := \left\{ \sum_{n=1}^N \alpha_n k(x_n, \cdot) \mid \alpha_n \in \mathbb{R},\, x_n \in \mathcal{X},\, N \in \mathbb{N} \right\} \tag{D.1}$$

with respect to norm induced by the inner product $\langle \sum_{n=1}^N \alpha_n k(x_n, \cdot), \sum_{n=1}^N \beta_n k(\widehat{x}_n, \cdot) \rangle := \sum_{n,n'=1}^N \alpha_n \beta_n k(x_n, \widehat{x}_{n'})$ for all $\alpha_n, \beta_n \in \mathbb{R}$, $x_n, \widehat{x}_n \in \mathcal{X}$ and $N \in \mathbb{N}$. The Hilbert space constructed in this way is an RKHS with kernel k [Wendland, 2004, Theorem 10.10].

Additional details about RKHS can be found in Berlinet and Thomas-Agnan [2004], Wendland [2004] and Steinwart and Christmann [2008]. Furthermore Schölkopf and Smola [2002] and Hofmann et al. [2008] describe how the theory of RKHS can be used in the context of machine learning.

D.1.2 Gaussian Random Elements and Gaussian Measures

Let $\big(\Omega, \mathcal{A}, \mathbb{P}\big)$ be the underlying physical probability space and $\big(H, \langle \cdot, \cdot \rangle\big)$ a (separable) Hilbert space.

The measurable mapping $F : \Omega \to H$ is called Gaussian Random Element (GRE) in H if $\langle F, h \rangle : \Omega \to \mathbb{R}$ is a Gaussian random variable for all $h \in H$ [1] [Van Neerven, 2008, Definition 4.1]. If F is a GRE then

[1] In this context a Gaussian random variable $X \sim \mathcal{N}(\mu, 0)$ is defined as Dirac measure at μ

$\mathbb{E}\big[||F||^2\big] < \infty$ [Van Neerven, 2008, Theorem 4.3] and we can define the mean element $\mathbb{E}[F] \in H$ and the covariance operator $\mathbb{C}[F] : H \to H$ via

$$\mathbb{E}[F] := \int F(\omega)\, d\mathbb{P}(\omega) \tag{D.2}$$

$$\mathbb{C}[F]h := \int F(\omega)\langle F, h\rangle\, d\mathbb{P}(\omega) - \langle m, h\rangle m \tag{D.3}$$

for $g \in H$, where the integrals are understood as Bochner integral. By standard rules for Bochner integrals, one obtains $\langle \mathbb{E}[F], h\rangle = \mathbb{E}\big[\langle F, h\rangle\big]$, $\langle \mathbb{C}[F]h, g\rangle = \mathbb{C}\big[\langle F, h\rangle, \langle F, g\rangle\big]$ and consequently $\langle F, h\rangle \sim \mathcal{N}\big(\langle \mathbb{E}[F], h\rangle, \langle \mathbb{C}[F]h, h\rangle\big)$. The covariance operator $\mathbb{C}[F] : H \to H$ is a positive, symmetric, trace class operator [Da Prato and Zabczyk, 2014, Proposition 2.16]. Conversely, for a given mean element $m \in H$ and given positive, symmetric, trace class operator $C : H \to H$ there exits a GRE $F : \Omega \to H$ such that $\mathbb{E}[F] = m$ and $\mathbb{C}[F] = C$ [Da Prato and Zabczyk, 2014, Proposition 2.18]. We write $F \sim \mathcal{N}(m, C)$ for a GRE with mean element m and covariance operator C.

A probability measure P defined on the Borel σ-algebra $\mathcal{B}$(H) is called Gaussian measure if and only the push-forward measure $T\#P : \mathcal{B}(\mathbb{R}) \to [0, 1], B \mapsto T\#P(B) := P\big(T^{-1}(B)\big)$ is a Gaussian measure on $\mathcal{B}(\mathbb{R})$ for all bounded, linear functionals $T : H \to \mathbb{R}$. Notice that by Riesz representation theorem every bounded, linear functional T is of the form $T = \langle \cdot, h\rangle$ for a $h \in H$. Hence, by definition, if $F : \Omega \to H$ is a GRE then the push-forward measure $F\#\mathbb{P}$ is a GM. Consequently, the statements about GREs outlined above carry over to GMs mutatis mutandis.

Bogachev [1998] is the authoritative source on Gaussian measures and discusses their properties on general Fréchet spaces. Gaussian measures on Banach and Hilbert spaces are described in Da Prato and Zabczyk [2014]. Van Neerven [2008] discusses Banach space valued Gaussian random elements.

D.2 Wasserstein Gradient Flow in $\mathcal{P}_2(H)$

Let $\big(H, \langle \cdot, \cdot\rangle\big)$ be a separable Hilbert space and

$$\mathcal{P}_2(H) := \left\{\mu : \mathcal{B}(H) \to [0, 1] \,\middle|\, \int_H ||u||^2\, d\mu(u) < \infty\right\} \tag{D.4}$$

be the space of Borel probability measures on $\mathcal{B}(H)$ with finite second moment. Here $\mathcal{B}(H)$ denotes the Borel σ-algebra on H. Define further for $Q \in \mathcal{P}_2(H)$ the Bochner spaces

$$L^2(Q) := \left\{f : H \to \mathbb{R} \,\middle|\, \int \big(f(u)\big)^2\, dQ(u) < \infty\right\} \tag{D.5}$$

$$L^2(Q;H) := \left\{ f : H \to H \,\middle|\, \int \langle f(u), f(u) \rangle \, dQ(u) < \infty \right\} \tag{D.6}$$

where functions are identified Q- almost everywhere.

We are interested in calculating the Wasserstein gradient flow for

$$L(Q) = \int_H \ell(u) \, dQ(u) + \mathrm{KL}(Q, \Pi), \tag{D.7}$$

$Q \in \mathcal{P}_2(H)$ where $\ell : H \to \mathbb{R}$ is a Fréchet differentiable function , $\Pi = \mathcal{N}(0, C)$ is a Gaussian measure on H with covariance operator $C : H \to H$ (cf. Appendix D.1.2) and KL the Kullback-Leibler divergence defined as

$$\mathrm{KL}(Q, \Pi) := \int \log \left(\frac{dQ}{d\Pi} \right) (u) \, dQ(u) \tag{D.8}$$

for Q dominated by Π where $dQ/d\Pi$ denotes the corresponding Radon-Nikodym derivative and $\mathrm{KL}(Q, \Pi) = \infty$ otherwise.

The first step in deriving the Wasserstein gradient flow is to calculate the Wasserstein gradient. We present a definition for the Wasserstein gradient below which is taken from Figalli and Glaudo [2021, Definition 4.2.2] adjusted to the Hilbert space case. A more general definition can be found in Chapter 11.1 of Ambrosio et al. [2005].

Definition D.2.1. Let $L : \mathcal{P}_2(H) \to [0, \infty]$ be a functional. The Wasserstein gradient at Q is the unique element $\phi \in L^2(Q; H)$ (if it exists) such that

$$\frac{d}{dt}\Big|_{t=0} L\big(Q(t)\big) = \int \langle \phi(u), v(u) \rangle \, Q(du) \tag{D.9}$$

for all smooth curves $\big(Q(t)\big)_{t \in (-\epsilon, \epsilon)} \subset \mathcal{P}_2(H)$ with $Q(0) = Q$ and *tangent vector* (explanation below) $v \in L^2(Q; H)$ at $t = 0$. We write $\nabla_W L[Q] : H \to H$ for the Wasserstein gradient ϕ at Q if it exists.

We will now discuss how to construct a curve $\big(Q(t)\big)_{t \in (-\epsilon, \epsilon)} \subset \mathcal{P}_2(H)$ at Q with tangent vector v. Let $u(t; u_0) \in H$ be the solution to the ODE

$$u(0) = u_0 \tag{D.10}$$

$$u'(t) = v\big(u(t)\big) \tag{D.11}$$

for $t \in (-\epsilon, \epsilon)$ with given initial value $u_0 \in H$. We assume that $v \in L^2(Q; H)$ is regular regular enough such that a solution exits for small enough $\epsilon > 0$. Then $\big(Q(t)\big)_{t \in (-\epsilon, \epsilon)} \subset \mathcal{P}_2(H)$ with $Q(t) := u(t; \cdot) \# Q$

is a smooth curve in $L^2(H)$ with tangent vector v at $Q \in \mathcal{P}_2(H)$.

Indeed, it suffices to show that (D.9) holds for all curves constructed via (D.10) and (D.11) [Ambrosio et al., 2005, Theorem 8.3.1].

Theorem D.2.2. *The Wasserstein gradient for L in (D.7) is given as*

$$\nabla_W L[Q](u) = D\ell(u) + D\log(dQ/d\Pi)(u) \tag{D.12}$$

for all $u \in H$ and $Q \in \mathcal{P}_2(H)$ dominated by $\Pi = \mathcal{N}(0, C)$. Notice that $L(Q) = \infty$ whenever Q is not dominated by Π and in this case the Wasserstein gradient is undefined.

Proof. Let $\big(Q(t)\big)$ be the curve constructed in (D.10) and (D.11).We denote by $q(t) := dQ(t)/d\Pi$ the Radon-Nikodym derivative of $Q(t)$ with respect to Π which exists for regular enough v in a small enough neighbourhood around $t = 0$. We sometimes write $q(t, u)$ instead of $q(t)(u)$ to avoid cluttered notation.Thank

We start the proof by deriving an equation for the time evolution of $q(t)$. The Hilbert space version of the Fokker-Planck equation (FPE) [Da Prato and Zabczyk, 2014, Section 14.2.2] states that

$$\frac{d}{dt}\int \varphi(u)\, dQ_t(u) = \int \big\langle v(u),\, D\varphi(u)\big\rangle\, dQ_t(u) \tag{D.13}$$

for all $\varphi \in C_b^2(H)$. For the LHS of the FPE we have

$$\frac{d}{dt}\int \varphi(u)\, dQ_t(u) = \int \varphi(u)\partial_t q(t,u)\, d\Pi(u) \tag{D.14}$$

and for the RHS of the FPE we obtain

$$\begin{aligned}
\int \big\langle v(u),\, D\varphi(u)\big\rangle\, dQ_t(u) &= \int \big\langle v(u),\, D\varphi(u)\big\rangle q(t,u)\, d\Pi(u) && \text{(D.15)}\\
&= \int \big\langle v(u)q(t,u),\, D\varphi(u)\big\rangle\, d\Pi(u) && \text{(D.16)}\\
&= \int -\mathrm{Tr}\left(D\big[v(u)q(t,u)\big]\right)\varphi(u)\, d\Pi(u) && \text{(D.17)}\\
&+ \int \big\langle C^{-1}u, v(u)q(t,u)\big\rangle\varphi(u)\, d\Pi(u) && \text{(D.18)}
\end{aligned}$$

where the last equality holds due to the integration by parts (IBP) formula for Gaussian measures on Hilbert spaces [Da Prato, 2006, Lemma 10.1 and Section 10.4] for regular enough v. Here Tr denotes the trace operator and all Fréchet derivatives are with respect to the u-variable. Technically speaking D is

the Friedrichs operator introduced in Section 10 of Da Prato [2006]. However, we will always assume Fréchet differentiability of all quantities involved and in this case D coincides with the Fréchet derivative. Combining both calculation gives the equality

$$\int \varphi(u)\partial_t q(t,u)\, d\Pi(u) \tag{D.19}$$

$$= \int -\mathrm{Tr}\left(D\big[v(u)q(t,u)\big]\right)\varphi(u)\, d\Pi(u) + \int \big\langle C^{-1}u, v(u)q(t,u)\big\rangle \varphi(u)\, d\Pi(u) \tag{D.20}$$

for all regular enough test functions $\varphi : H \to \mathbb{R}$ and consequently

$$\partial_t q(t,u) = -\mathrm{Tr}\left(D\big[v(u)q(t,u)\big]\right) + \big\langle C^{-1}u, v(u)q(t,u)\big\rangle \tag{D.21}$$

holds for all $u \in H$ and all t in a small enough interval around $t = 0$. Also note that $q(0,\cdot) = q := dQ/d\Pi$ by construction.

We will now calculate the Wasserstein gradient for L at Q. By standard rules for Radon-Nikodym derivatives we obtain

$$L\big(Q(t)\big) = \underbrace{\int \ell(u)q(t,u)\, d\Pi(u)}_{(a)} + \underbrace{\int \log q_t(u) q_t(u)\, d\Pi(u)}_{(b)}. \tag{D.22}$$

We calculate for (a)

$$\frac{d}{dt}\int \ell(u)q(t,u)\, d\Pi(u) = \int \ell(u)\partial_t q(t,u)\, d\Pi(u) \tag{D.23}$$

$$= \int \ell(u)\left(-\mathrm{Tr}\left(D\big[v(u)q(t,u)\big]\right) + \big\langle C^{-1}u, v(u)q(t,u)\big\rangle\right) d\Pi(u) \tag{D.24}$$

$$= \int \big\langle D\ell(u), v(u)q(t,u)\big\rangle\, d\Pi(u) \tag{D.25}$$

$$= \int \big\langle D\ell(u), v(u)\big\rangle\, dQ_t(u), \tag{D.26}$$

where the second equality follows from (D.21) and the third euality from IBP. For (b) we calculate

$$\frac{d}{dt}\int \log q_t(u) q_t(u)\, d\Pi(u) \tag{D.27}$$

$$= \int \partial_t q(t,u)(1 + \log q_t(u))\, d\Pi(u) \tag{D.28}$$

$$= \int \left(-\mathrm{Tr}\left(D\big[v(u)q(t,u)\big]\right) + \big\langle C^{-1}u, v(u)q(t,u)\big\rangle\right)\big(1 + \log q_t(u)\big)\, d\Pi(u) \tag{D.29}$$

$$= \int \big\langle D(1 + \log q_t(u)), v(u)q(t,u)\big\rangle\, d\Pi(u) \tag{D.30}$$

$$= \int \big\langle D(\log q_t(u)), v(u) \big\rangle \, dQ_t(u) \tag{D.31}$$

where we again use (D.21) and IBP. We put the calculation for (a) and (b) together and obtain

$$\frac{d}{dt} L\big(Q(t)\big) = \int \big\langle D\ell(u) + D \log q_t(u), v(u) \big\rangle \, d\Pi(u). \tag{D.32}$$

We evaluate the RHS of (D.32) for $t = 0$ and see that the Wasserstein gradient at Q is given as

$$\nabla_W L[Q](u) = D\ell(u) + D \log q(u) \tag{D.33}$$

for $u \in H$ with $q = dQ/d\Pi$. Note that this is precisely what we would expect from the finite-dimensional case. □

The next step is to identify a suitable stochastic process $\big(F(t)\big)$ with $F(t) : \Omega \to H$ such that $\text{Law}[F(t)] = Q(t)$ where $Q(t)$ is the WGF at time t.

Theorem D.2.3. *Let $\big(F(t)\big)_{t\in[0,T]}$ be the solution (which we assume exists, see Hairer et al. [2007b] or Da Prato and Zabczyk [2014] for conditions) to*

$$F(0) \sim Q_0 \tag{D.34}$$

$$dF(t) = -\big(D\ell\big(F(t)\big) + C^{-1}F(t)\big)\, dt + \sqrt{2} dW(t) \tag{D.35}$$

for $t \in [0, T]$ where $Q_0 \in \mathcal{P}_2(H)$ is given, C^{-1} is the inverse of the covariance operator C and $\big(W(t)\big)$ a cylindrical Wiener process. Then $Q(t) := \text{Law}\big[F(t)\big]$ follows the Wasserstein gradient flow for L in (D.9) and starts at Q_0.

Proof. Recall that the loss L in (D.9) for all Q dominated by Π can be written as

$$L(Q) = \text{KL}(Q, \Pi^*) + \text{ const.} \tag{D.36}$$

where Π^* is the Bayesian posterior [Wild and Wynne, 2022]. Recall further that by Bayes theorem

$$\frac{d\Pi^*}{d\Pi}(u) = \frac{p(y|u)}{p(y)} \tag{D.37}$$

for all $u \in H$ where $p(y|u)$ is the likelihood function, $p(y) = \int p(y|u)\, d\Pi(u)$ the marginal likelihood and $\ell(u) := -\log p(y|u)$ as introduced in the main text. Note that Π^* is log-concave in the sense of Definition 9.4.9 in Ambrosio et al. [2005] as long as $\ell(u)$ is convex [Ambrosio et al., 2005, Theorem 9.4.11]. We

will assume ℓ to be convex from now on which is typically true for functional losses. Theorem 11.2.12 in Ambrosio et al. [2005] implies that the Wasserstein gradient flow $\big(Q(t)\big)$ for L exists and further that the Radon-Nikodym derivative of $Q(t)$ with respect to Π^* exists and that $\rho_t := dQ(t)/d\Pi^*$ satisfies

$$\int_0^T \int_H \partial_t \varphi(t,u) - \langle D \log \rho_t(u), D\varphi(t,u) \rangle \, dQ_t(u) \, dt = 0 \tag{D.38}$$

for all test functions $\varphi : [0,T] \times H \to \mathbb{R}$. We want to rewrite this equation in terms of $q(t) := dQ(t)/d\Pi$. First note that $q(t)$ exists since Π and Π^* are equivalent in our case [Ghosal and van der Vaart, 2017, Section 1.3]. We further have by the chain-rule for Radon-Nikodym densities

$$\begin{aligned}
\log \rho_t(u) &= \log(dQ(t)/d\Pi^*)(u) && \text{(D.39)}\\
&= \log q_t(u) + \log(d\Pi/d\Pi^*)(u) && \text{(D.40)}\\
&= \log q_t(u) - \log(d\Pi^*/d\Pi)(u) && \text{(D.41)}\\
&= \log q_t(u) - \log p(y|u) + \log p(y) && \text{(D.42)}\\
&= \log q_t(u) + \ell(u) + \log p(y). && \text{(D.43)}
\end{aligned}$$

We plug this into (D.38) and obtain the dynamics for $q(t)$ as

$$\int_0^T \int_H \partial_t \varphi(t,u) - \langle D \log q_t(u) + D\ell(u), D\varphi(t,u) \rangle \, dQ_t(u) \, dt = 0. \tag{D.44}$$

Since $D \log q_t + D\ell = \nabla_W L[Q(t)]$ (cf. Theorem D.2.2) we conclude that $q(t)$ indeed satisfies (6.2) and therefore follows the WGF.

However note that every solution to (D.44) satisfies (differentiate w.r.t. to time)[2]

$$\frac{d}{dt} \int \psi(u) dQ_t(u) = - \int_H \langle D \log q_t(u) + D\ell(u), D\psi(u) \rangle \, dQ_t(u) \tag{D.45}$$

for all t and all test functions $\psi : H \to \mathbb{R}$. The RHS of (D.45) is equal to

$$\begin{aligned}
&- \int_H \langle D \log q_t(u) + D\ell(u), D\psi(u) \rangle \, dQ_t(u) && \text{(D.46)}\\
&= - \int \langle D\ell(u), D\psi(u) \rangle dQ_t(u) - \int_H \langle D \log q_t(u), D\psi(u) \rangle \, dQ_t(u) && \text{(D.47)}\\
&= - \int \langle D\ell(u), D\psi(u) \rangle dQ_t(u) - \int_H \langle Dq_t(u), D\psi(u) \rangle \, d\Pi(u) && \text{(D.48)}\\
&= - \int \langle D\ell(u), D\psi(u) \rangle dQ_t(u) + \int \mathrm{Tr}\big[D^2\psi(u)\big] \, dQ_t(u) - \int \langle C^{-1}u, D\psi(u) \rangle dQ_t(u) && \text{(D.49)}
\end{aligned}$$

[2]The reverse may be easier: Multiply both sides of (D.45) with a test function and then integrate with respect to time. This gives (D.44).

where the last equality follows from integration by parts for Gaussian measures on Hilbert spaces [Da Prato and Zabczyk, 2002, Lemma 11.1.9]. We combine (D.45) and (D.49) to obtain

$$\frac{d}{dt}\int \psi(u)dQ_t(u) = \int \mathcal{A}_t\psi(u)dQ_t(u) \tag{D.50}$$

where $\mathcal{A}_t$ is the Kolmogorov operator defined as

$$\mathcal{A}_t\psi(u) = \mathrm{Tr}\big[D^2\psi(u)\big] + \big\langle -D\ell(u) - C^{-1}u, D\psi(u)\big\rangle. \tag{D.51}$$

We recognise (D.50) as the Fokker-Planck equation (FPE) associated with the solution to (D.35) (see Chapter 14.2.2 of Da Prato and Zabczyk [2014] or Bogachev et al. [2008,0] for earlier references). Note that any solution $\big(Q(t)\big)$ to the FPE solves (D.44) by reversing the above argument. As a consequence $\big(Q(t)\big)$ satisfies the WGF if and only if it is a solution to the FPE. We can therefore simulate the SDE (D.35) in order to follow the WGF. □

D.3 The Moments of the Posterior Measure

Our Bayesian model consist of:

- A prior measure $\Pi \in \mathcal{P}_2(H)$
- A likelihood function $p : \mathbb{R}^N \times H \to [0,\infty)$, i.e. p is $\mathcal{B}(\mathbb{R}^N) \otimes \mathcal{B}(H)$- $\mathcal{B}([0,\infty))$ measurable and there exists a measure $\mu \in \mathcal{P}(\mathbb{R}^N)$ such that

$$\int p(y|f)\, d\mu(y) = 1. \tag{D.52}$$

According to Bayes' theorem, the posterior Π^* exists and has a Radon-Nikodym derivative with respect to the prior Π given as

$$\frac{d\Pi^*}{d\Pi}(f) = \frac{p(y|f)}{p(y)}, \tag{D.53}$$

where $p(y) := \int p(y|f)\, d\Pi(f)$. By definition of the likelihood function, we know that $\gamma(A) := \int_A p(y)\, d\mu(y)$, $A \in \mathcal{B}(\mathbb{R}^N)$, is a probability measure. We can now state the theorem.

Theorem D.3.1. *The Bayesian posterior satisfies $\Pi^* \in \mathcal{P}_2(H)$ for γ-almost every $y \in \mathbb{R}^N$.*

Proof. Let $\big(\Omega, \mathcal{A}, \mathbb{P}\big)$ be an (sufficiently large) underlying probability space and $F : \Omega \to H$, $Y : \Omega \to$

$\mathbb{R}^N$ two measurable mappings such that $F\#\mathbb{P} = \Pi$, $Y\#\mathbb{P} \sim \gamma$ and

$$\mathbb{P}(Y \in A|F = f) = \int_A p(y|f)\, d\mu(y) \tag{D.54}$$

for all $A \in \mathcal{B}(\mathbb{R}^N)$, $f \in H$. By the above construction, the claim of Theorem D.3.1 holds if and only if

$$\mathbb{E}\big[||F||^2|Y = y\big] < \infty \tag{D.55}$$

for γ-almost every $y \in \mathbb{R}^N$. However, we know by the tower-property of expected values, that

$$\mathbb{E}\big[||F||^2\big] = \int \mathbb{E}\big[||F||^2|Y = y\big]\, d\gamma(y). \tag{D.56}$$

Since $\mathbb{E}[||F||^2] < \infty$ by Fernique's theorem [Da Prato and Zabczyk, 2014, Theorem 2.7], we can immediately conclude that $\mathbb{E}\big[||F||^2|Y = y\big] < \infty$ for γ-almost every $y \in \mathbb{R}^N$. □

Remark D.3.2. The proof above is identical for finite-dimensional parameters. However, we felt that including an infinite-dimensional version may still provide some additional benefit.

Furthermore, the result can easily be generalised to arbitrary moments, in the sense that, for a fixed $l \in \mathbb{N}$, it holds that $\mathbb{E}\big[||F||^l\big] < \infty$ implies $\mathbb{E}\big[||F||^l \,|\, Y = y\big] < \infty$ for γ-almost every $y \in \mathbb{R}^N$.

D.4 Gaussian Random Elements with Values in the RKHS

In this section we prove the existence of a GRE $F : \Omega \to H_k$ with covariance operator $C : H_k \to H_k$ defined in (6.9). This allows us to effectively parameterise RKHS valued GREs via the kernel k from the RKHS $H = H_k$.

Lemma D.4.1. *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive, symmetric and continuous kernel on a compact metric space $\mathcal{X}$ and $H = H_k$ the associated RKHS. Let further $\nu \in \mathcal{P}(\mathcal{X})$ be a Borel probability measure with full support and assume that $\int k(x,x)\, d\nu(x) < \infty$. Then there exists a Gaussian random element $F \sim \mathcal{N}(0, C)$ in H with covariance operator $C : H \to H$ given as*

$$Cf = \int k(\cdot, x') f(x')\, d\nu(x'). \tag{D.57}$$

Proof. Let

$$L^2(\nu, \mathbb{R}) := \left\{ f : \mathcal{X} \to \mathbb{R} \,:\, \int \big(f(x)\big)^2\, d\nu(x) < \infty \right\} \tag{D.58}$$

be the space of equivalence classes of ν-almost everywhere identical square-integrable functions with inner product denoted by $\langle \cdot, \cdot \rangle_2$. Define $T_k : L^2(\nu, \mathbb{R}) \to L^2(\nu, \mathbb{R})$´as $T_k f = \int k(\cdot, x') f(x') \, d\nu(x')$. Note that T_k and C have the same functional form but are defined on different spaces.

Under the assumptions on k we know that T_k is self-adjoint, compact operator and therefore the spectral theorem guarantees the existence of an orthonormal basis $\{b_n\}_{n=1}^{\infty} \subset L_2(\nu, \mathbb{R})$ and eigenvalues $\{\lambda_n\}_{n=1}^{\infty} \subset \mathbb{R}_+$ such that

$$T_k f = \sum_{n=1}^{\infty} \lambda_n \langle f, b_n \rangle_2 \, b_n \tag{D.59}$$

where the sum converges in $L^2(\nu, \mathbb{R})$. We can now define $S : \mathcal{L}^2(\nu, \mathbb{R}) \to H_k$ via

$$Sf = \sum_{n=1}^{\infty} \sqrt{\lambda_n} \langle f, b_n \rangle_2 \, b_n. \tag{D.60}$$

It is well-known that S an isometric isomorphism between $L^2(\nu, \mathbb{R})$ and H_k [Steinwart and Christmann, 2008, Theorem 4.5.1]. In particular this means

$$\langle Sf, Sg \rangle = \langle f, g \rangle_2 \tag{D.61}$$

for all $f, g \in L^2(\nu, \mathbb{R})$.

According to Theorem 1 in Wild et al. [2022], there exists a GRE G in $L^2(\nu, \mathbb{R})$ with $G \sim \mathcal{N}(0, T_k)$. Our goal is now to show that $F := S_k \circ G$ is a Gaussian random element in H_k with covariance operator C.

Lemma 5.6 in Kukush [2020] implies that F is a GRE in H_k with covariance operator ST_kS^* where S^* is the adjoint operator of S. It remains to show that $C = ST_kS^*$.

First, note that $S^* : H_k \to \mathcal{L}^2(\nu, \mathbb{R})$ is characterised by satisfying

$$\langle Sb_i, b_j \rangle = \langle b_i, S^* b_j \rangle_2 \tag{D.62}$$

for all $(i, j) \in \mathbb{N}^2$ since $\{b_n\}_{n=1}^{\infty}$ is an orthonormal system in H_k. By construction, we know that $Sb_j = \sqrt{\lambda_j} b_j$ and therefore

$$\langle Sb_i, b_j \rangle = \frac{1}{\sqrt{\lambda_j}} \langle Sb_i, Sb_j \rangle = \frac{1}{\sqrt{\lambda_j}} \langle b_i, b_j \rangle_2 \tag{D.63}$$

where the last equality follows from the isometry property of S. Combining (D.62) and (D.63) leads to

$$\langle b_i, S^* b_j \rangle_2 = \frac{1}{\sqrt{\lambda_j}} \langle b_i, b_j \rangle_2 \tag{D.64}$$

for all $(i, j) \in \mathbb{N}^2$ and therefore $S^* b_j = \frac{1}{\sqrt{\lambda_j}} b_j$ for all $j \in \mathbb{N}$. It now follows immediately that

$$(S T_k S^*) b_j = \frac{1}{\sqrt{\lambda_j}} S T_k b_j = \lambda_j b_j. \tag{D.65}$$

On the other hand, we have

$$C b_j = T_k b_j = \lambda_j b_j \tag{D.66}$$

for all $j \in \mathbb{N}$ by the spectral theorem and therefore $C = S T_k S^*$ by density of $\{b_n\}_{n=1}^{\infty}$ in H_k as claimed. □

We show that F naturally induces a Gaussian process G albeit with a new kernel r which is a smoothed version of k.

Lemma D.4.2. *Let $F \sim \mathcal{N}(0, C)$ be the GRE in the RKHS $H = H_k$ with covariance operator C defined in (D.57) (cf. Lemma D.4.1). Define $G(x) := \langle F, k(x, \cdot) \rangle$ and $G := \{G(x) : x \in \mathcal{X}\}$. Then G is a Gaussian process with kernel r given by*

$$r(x, x') = \int k(x, z) k(z, x') \, d\nu(z) \tag{D.67}$$

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

Proof. The process G is a GP with kernel r if and only if $G(X) \sim \mathcal{N}\big(0, r(X, X)\big)$ for all $X = (x_1, \ldots, x_N) \in \mathcal{X}^N$. The later is by definition equivalent to

$$\alpha^T G(X) \sim \mathcal{N}(0, \alpha^T r(X, X) \alpha). \tag{D.68}$$

for all $\alpha \in \mathbb{R}^N$ which we will prove hereafter. Let now $\alpha \in \mathbb{R}^N$ and $X = (x_1, \ldots, x_N) \in \mathcal{X}^N$ arbitrary. We then have

$$\alpha^T G(X) = \sum_{n=1}^{N} \alpha_n \langle G, k(x_n, \cdot) \rangle \tag{D.69}$$

$$= \big\langle G, \underbrace{\sum_{n=1}^{N} \alpha_n k(x_n, \cdot)}_{=:h} \big\rangle \tag{D.70}$$

where the first equality follows from the reproducing property and the second by bilinearity of the scalar product. By definition of GREs (cf. Appendix D.1.2), we infer from (D.70) that $\alpha^T G(X)$ is Gaussian random variable and it's variance is given as

$$\mathbb{V}[\alpha^T G(X)] = \langle Ch, h \rangle = \sum_{n,n'=1}^{N} \alpha_n \alpha_{n'} \langle Ck(x_n, \cdot), k(x_{n'}, \cdot) \rangle. \tag{D.71}$$

By standard rules for Bochner integrals, we further obtain

$$\begin{aligned}
\langle Ck(x_n, \cdot), k(x_{n'}, \cdot) \rangle &= \left\langle \int k(\cdot, z) k(z, x_n) \, d\nu(z), k(x_{n'}, \cdot) \right\rangle && \text{(D.72)} \\
&= \int \langle k(\cdot, z), k(x_{n'}, \cdot) \rangle k(z, x_n) \, d\nu(z) && \text{(D.73)} \\
&= \int k(z, x_{n'}) k(z, x_n) \, d\nu(z) && \text{(D.74)} \\
&= r(x_n, x_{n'}) && \text{(D.75)}
\end{aligned}$$

for all $n, n' = 1, \ldots, N$ which proves the claim □

Remark D.4.3. This kernel r was already proposed in in Section 3.1 of Filippi et al. [2016] in an effort to construct a Gaussian processes with sample paths that lie in the RKHS $H = H_k$. Our derivation gives a natural interpretation of the corresponding GP as being derived from a GP with sample paths in $L^2(\nu)$ which is equipped with additional smoothness by virtue of the isometric isomorphism S in (D.60).

D.5 The Influence of the Measure ν

The Gaussian process G constructed in Lemma D.4.2 has kernel r given as

$$r(x, x') = \int k(x, s) k(s, x') \, d\nu(s) \tag{D.76}$$

for all $x, x' \in \mathcal{X}$. In principle any Borel measure $\nu \in \mathcal{P}(\mathcal{X})$[3] which satisfies the conditions in Lemma D.4.1 leads to a valid GRE in H_k.

One idea is therefore to choose a combination of k and ν such that r can be calculated analytically. This strategy was proposed previously [Cialenco et al., 2012, Filippi et al., 2016] and can be very effective. However, it severely restricts the class of available kernels k since few kernel exists that lead to a tractable expression in (D.76).

[3] In fact ν is only required to be finite and does not need to be a probability measure.

Alternatively, we can choose a probability measure $\nu \in \mathcal{P}(\mathcal{X})$ from which which it is easy to generate samples $\widehat{X}_1, \dots \widehat{X}_{N_S} \sim \nu$ where $N_S \in \mathbb{N}$ is the number of samples. This leads to the Monte Carlo estimator

$$r(x, x') \approx \frac{1}{N_S} k(x, \widehat{X}) k(\widehat{X}, x') := \frac{1}{N_S} \sum_{n=1}^{N_S} k(x, \widehat{x}_n) k(\widehat{x}_n, x') \tag{D.77}$$

for r where $\widehat{X} := (\widehat{X}_1, \dots \widehat{X}_{N_S})$.

We now want to illustrate that the kernel effectively becomes useless, if the input data distribution, i.e. the law of $X_1, \dots, X_N$ deviates strongly from ν. To this end, assume that we use a squared exponential kernel k with

$$k(x, x') = \exp\big(- |x - x'|^2 \big) \tag{D.78}$$

for $x, x' \in \mathbb{R}$. Assume further that the data is given as $X_1, \dots, X_N \sim \mathcal{N}(0, 1)$. We now choose $\nu = \mathcal{N}(10, 1)$ and generate samples $\widehat{X}_1, \dots, \widehat{X}_{N_S} \sim \nu$. Take now two points x, x' which are in a high-density region of $\mathcal{N}(0, 1)$, e.g. $x = 0$ and $x' = 0.5$, then

$$r(x, x') = \mathbb{E}_\xi \Big[\exp\big(- |\xi|^2 - |0.5 - \xi|^2 \big) \Big] \approx 1.3 \cdot 10^{-43} \tag{D.79}$$

where $\xi \sim \nu = \mathcal{N}(10, 1)$. Put differently, a kernel matrix based on r would be extremely uninformative. This is why we choose $\nu = \sum_{n=1}^N \delta_{x_N}$ or an approximation based on sub-samples of $x_1, \dots, x_N$ in all our experiments.

D.6 Langevin SDE in ONB Representation

The time evolution of the SDE in H is given as (cf. Theorem D.2.3)

$$dF(t) = - \big(D\ell\big(F(t)\big) + C^{-1} F(t) \big)\, dt + \sqrt{2} dW(t) \tag{D.80}$$

for $t \in (0, T]$ and $F(0) = F_0$ with $F_0 \in H$ given. Define now $F^m(t) := \langle F(t), e_m \rangle$ where $\{\lambda_m, e_m\}_{m=1}^\infty$ is the spectral decomposition of the covariance operator C in (6.9). Define $\phi : H_k \to \mathbb{R}$ with $\phi(f) := \langle f, e_m \rangle$ and by linearity it's Fréchet derivatives are given as $D\phi(f) = e_m$ and $D^2\phi = 0$ and consequently by Ito's rule [Da Prato and Zabczyk, 2014, Chapter 4.4] we obtain

$$dF^m(t) = d\phi\big(F(t)\big) \tag{D.81}$$

$$= -\langle D\ell\big(F(t)\big) + C^{-1}F(t), e_m\rangle dt + \sqrt{2}d\langle W(t), e_m\rangle \tag{D.82}$$

$$= -\langle D\ell\big(F(t)\big), e_m\rangle dt - \langle C^{-1}F(t), e_m\rangle dt + \sqrt{2}d\langle W(t), e_m\rangle. \tag{D.83}$$

Recall, that our loss ℓ is of the form (6.6) and we therefore can simplify further as

$$dF^m(t) = -\sum_{n=1}^{N}(\partial_2 c)\big(y_n, F_t(x_n)\big)\langle k(x_n,\cdot), e_m\rangle dt - \langle F(t), C^{-1}e_m\rangle dt + \sqrt{2}d\langle W(t), e_m\rangle \tag{D.84}$$

$$= -\Big(\sum_{n=1}^{N}(\partial_2 c)(y_n, F_t(x_n))e_m(x_n) + \frac{F^m(t)}{\lambda_m}\Big)dt + \sqrt{2}dB^m(t), \tag{D.85}$$

where we used the reproducing property and that C^{-1} is self adjoint. Note that $B^m(t) := \langle W(t), e_m\rangle$ $(m = 1, \dots, M)$ are stochastically independent Brownian motions by standard properties of the cylindrical Wiener process [Da Prato and Zabczyk, 2014, Chapter 4.1.2].

D.7 The Nyström Method

The Nyström method [Williams and Seeger, 2001] is designed to approximate eigenvalue-eigenfunction pairs $\{\lambda_n, b_n\}_{n=1}^{\infty}$ of the kernel integral operator $T_k : \mathcal{L}_2(\nu;\mathbb{R}) \to \mathcal{L}_2(\nu;\mathbb{R})$ defined as

$$T_k f := \int k(\cdot, x')f(x')\, d\nu(x') \tag{D.86}$$

which we introduced in Appendix D.4. For $X_1, \dots, X_N$ independent samples from ν, the Nyström method calculates the spectral decomposition of $\frac{1}{N}k(x_{1:N}, x_{1:N}) := \frac{1}{N}\big(k(x_n, x_{n'})\big)_{n,n'=1}^{N} \in \mathbb{R}^{N\times N}$ as

$$\frac{1}{N}k(x_{1:N}, x_{1:N}) = V\widehat{\Lambda}V^T \tag{D.87}$$

where $V = (v_1|\dots|v_N)$ and $\widehat{\Lambda} := \operatorname{diag}(\widehat{\lambda}_1, \dots, \widehat{\lambda}_N)$. Then

$$\widehat{b}_n(x) := \frac{1}{\sqrt{N\widehat{\lambda}_n}}v_n^T k(X, x) := \frac{1}{\sqrt{N\widehat{\lambda}_n}}\sum_{j=1}^{N} v_{nj}k(x_j, x)$$

approximates $b_n(x)$ and $\widehat{\lambda}_m$ approximates λ_m [Williams and Seeger, 2001].

However, C has the same eigenvalues as T_k and the eigenfunctions are related [Steinwart and Christmann, 2008, Theorem 4.51] with $e_n = \sqrt{\lambda_n}b_n$. We can therefore use

$$\widehat{e}_n(x) := \sqrt{\widehat{\lambda}_n}\widehat{b}_n(x) = \frac{1}{\sqrt{N\widehat{\lambda}_n}}v_n^T k(X, x) \tag{D.88}$$

as approximation for e_n.

In many cases, using all available data $X_1, \dots, X_N$ would be computationally prohibitive since the spectral decomposition in (D.87) costs $\mathcal{O}(N^3)$. However, for typical kernels, such as the squared exponential, the eigenvalues decay very rapidly and only the first few eigenvalues contribute meaningfully to explain the data. We can therefore chose a subsample $z_1, \dots, z_M$ of $x_1, \dots, x_N$ and apply the Nyström method for $z_{1:M} := (z_1, \dots, z_M)$ where $M \ll N$ which leads to $\mathcal{O}(M^3)$ costs. We describe a heuristic in Appendix D.14 for choosing $z_{1:M}$.

D.8 Sufficiency of Nyström Projections

Let $F \sim \mathcal{N}(0, C)$ be a GRE and $Y_{1:N} \sim \gamma$ be two random mappings such that (cf. Appendix D.3 for a construction)

$$\mathbb{P}(Y_{1:N} \in A | F = f) = \int_A p(y_{1:N}|f)\, d\mu(y_{1:N}). \tag{D.89}$$

We assume that the negative log-likelihood is of the form

$$-\log p(y_{1:N}|f) = \sum_{n=1}^{N} c(y_n, f(x_n)) + \text{ const.} \tag{D.90}$$

We define $\psi : \mathbb{R}^N \to \mathbb{R}$ as $\psi\big(f(x_{1:N})\big) := \exp(-\sum_{n=1}^N c(y_n, f(x_n) - \text{const.})$. Then it holds that

$$\mathbb{P}\big(F(x^*) \in B \,|\, y_{1:N}, \widehat{F}^{1:N}\big) = \mathbb{P}\big(F(x^*) \in B \,|\, \widehat{F}^{1:N}\big) \tag{D.91}$$

for all $x^* \in \mathcal{X}^{N_*}$ and Borel sets $B \subset \mathbb{R}^{N_*}$. Furthermore, if $\nu = \frac{1}{N}\sum_{n=1}^N \delta_{x_n}$, then we have

$$\widehat{\tau}_\infty = \text{Law}\big[\widehat{F}^{1:N} | y_{1:N}\big] = \frac{1}{Z}\int_{(\cdot)} \exp(-\widehat{V}_\infty(u))\, du \tag{D.92}$$

where $Z := \int \exp(-\widehat{V}_\infty(u))\, du$ and $\widehat{V}_\infty : \mathbb{R}^N \to \mathbb{R}$ given as

$$\widehat{V}_\infty(u) := \sum_{n=1}^{N} c\big(y_n, \sum_{m=1}^{N} \widehat{e}_m(x_n) u_m\big) + \frac{1}{2} u^T \widehat{\Lambda}^{-1} u. \tag{D.93}$$

with $\widehat{\Lambda}$ defined in Appendix D.7.

Proof. Recall that $\widehat{F}^{1:N} := \big(\langle F, \widehat{e}_1\rangle, \dots, \langle F, \widehat{e}_N\rangle\big)^T$ with

$$\widehat{e}_n(x) = \widehat{v}_n^T k_X(x) = \sum_{i=1}^N \widehat{v}_{ni} k(x, x_i), \tag{D.94}$$

where $\widehat{v}_n = \frac{1}{\sqrt{N\widehat{\lambda}_n}} v_n \in \mathbb{R}^N$ (cf. Appendix D.7). We now want to show that the orthogonal projection $\widehat{\Pr}[f] = \sum_{n=1}^N \langle f, \widehat{e}_n\rangle \widehat{e}_n$ is exact for all data points $x_1, \dots, x_N$ which were used in the Nyström method, i.e $\widehat{\Pr}[f](x_n) = x_n$ for all $n = 1, \dots, N$. To this end, we define $k_X(\cdot) = \big(k(x_1, \cdot), \dots, k(x_N, \cdot)\big)^T$ and calculate

$$\begin{aligned}
\widehat{\Pr}[f](x_n) &= \sum_{i=1}^N \langle f, \widehat{e}_i\rangle \widehat{e}_i(x_n) && \text{(D.95)}\\
&= \sum_{i=1}^N \widehat{v}_i^T f(x_{1:N}) \widehat{v}_i^T k_X(x_n) && \text{(D.96)}\\
&= f(x_{1:N})^T \Big(\sum_{i=1}^N \widehat{v}_i \widehat{v}_i^T \Big) k_X(x_n) && \text{(D.97)}\\
&= f(x_{1:N})^T \Big(\frac{1}{N} \sum_{i=1}^N \frac{1}{\lambda_i} v_i v_i^T \Big) k_X(x_n) && \text{(D.98)}\\
&= f(x_{1:N})^T k(x_{1:N}, x_{1:N})^{-1} k_X(x_n) && \text{(D.99)}\\
&= f(x_n), && \text{(D.100)}
\end{aligned}$$

for all $n = 1, \dots, N$. This can equivalently be expressed as

$$f(x_{1:N}) = \widehat{\Pr}[f](x_{1:N}) = \sum_{i=1}^N \underbrace{\langle f, \widehat{e}_i\rangle}_{=:\widehat{f}^i} \widehat{e}_i(x_{1:N}) \tag{D.101}$$

for all $f \in H_k$. Consequently, we have

$$p(y_{1:N}|f) = \psi\Big(\sum_{i=1}^N \widehat{f}^i \widehat{e}_i(x_{1:N}) \Big) \tag{D.102}$$

and conclude that the conditional density of $Y_{1:N}$ given $F = f$ is a measurable function of $\widehat{f}^{1:N}$. We now write $p(f(x^*)|\widehat{f}^{1:N}, y_{1:N})$ for the pdf of $F(x^*)$ given $(\widehat{F}^{1:N}, Y_{1:N}) = (\widehat{f}^{1:N}, y_{1:N})$ and similarly for other marginal and conditional distributions involved and obtain

$$p\big(f(x^*)|\widehat{f}^{1:N}, y\big) = \frac{p\big(f(x^*), \widehat{f}^{1:N}, y_{1:N}\big)}{p\big(\widehat{f}^{1:N}, y\big)} \tag{D.103}$$

$$= \frac{p\big(y_{1:N}|f(x^*), \widehat{f}^{1:N}\big)p\big(f(x^*)|\widehat{f}^{1:N}\big)p\big(\widehat{f}^{1:N}\big)}{p\big(y_{1:N}|\widehat{f}^{1:N}\big)p\big(\widehat{f}^{1:N}\big)}. \tag{D.104}$$

Here, we notice that $p\big(y_{1:N}|f(x^*), \widehat{f}^{1:N}\big) = p(y_{1:N}|\widehat{f}^{1:N})$ because $p(y_{1:N}|f)$ is a measurable function of $\widehat{f}^{1:N}$ and consequently we obtain

$$p\big(f(x^*)|\widehat{f}^{1:N}, y_{1:N}\big) = p(f(x^*)|\widehat{f}^{1:N}) \tag{D.105}$$

which shows (D.91). It remains to show (D.92). However, analogous to what we do in in Appendix D.10, it is easy to show that (6.14) is a Langevin diffusion with potential

$$\widehat{V}_\infty(u) := \sum_{n=1}^N c\big(y_n, \sum_{m=1}^N \widehat{e}_m(x_n)u_m\big) + \frac{1}{2}u^T\widehat{\Lambda}^{-1}u. \tag{D.106}$$

Consequently, we immediately infer that $\widehat{\tau}_\infty \propto \exp(-\widehat{V}_\infty)$ which proves the second equality in (D.92). Furthermore, it follow from (D.102) that

$$-\log p(y|\widehat{f}^{1:N}) = -\log \psi\big(\sum_{m=1}^N \widehat{f}^m\widehat{e}_m(x_{1:N})\big) \tag{D.107}$$

$$= \sum_{n=1}^N c\big(y_n, \sum_{m=1}^N \widehat{e}_m(x_n)\widehat{f}^m\big) \tag{D.108}$$

and for the prior we know that $\widehat{F}^{1:N} \sim \mathcal{N}\Big(0, \big(\langle C\widehat{e}_m, \widehat{e}_{m'}\rangle\big)_{m,m'=1}^N\Big)$ by definition of a GRE. By definition, v_m is the eigenvector of $\frac{1}{N}k(x_{1:N}, x_{1:N})$ with eigenvalue $\widehat{\lambda}_m$ and further $r(x_{1:N}, x_{1:N}) = \frac{1}{N}k(x_{1:N}, x_{1:N})^2$ due to $\nu = \sum_{n=1}^N \delta_{x_n}$ which leads to

$$\langle C\widehat{e}_m, \widehat{e}_{m'}\rangle = \widehat{v}_m^T r(x_{1:N}, x_{1:N})\widehat{v}_{m'} \tag{D.109}$$

$$= \frac{1}{N}\widehat{v}_m^T\big(k(x_{1:N}, x_{1:N})\big)^2\widehat{v}_{m'} \tag{D.110}$$

$$= \frac{1}{\sqrt{\widehat{\lambda}_m\widehat{\lambda}_{m'}}}v_m^T\big(\frac{1}{N}k(x_{1:N}, x_{1:N})\big)^2 v_{m'} \tag{D.111}$$

$$= \widehat{\lambda}_m\delta_{m,m'}, \tag{D.112}$$

where $\delta_{m,m'}$ denotes the Kronecker delta. As a result, we have by Bayes theorem

$$p_{\widehat{F}^{1:N}|Y_{1:N}}(u|y_{1:N}) \propto \exp\Big(\log p(y_{1:N}|\widehat{F}^{1:N} = u) + \log p_{\widehat{F}^{1:N}}(u)\Big) \tag{D.113}$$

$$= \exp\Big(-\sum_{n=1}^N c\big(y_n, \sum_{m=1}^N \widehat{e}_m(x_n)u_m\big) - \frac{1}{2}u^T\widehat{\Lambda}^{-1}u\Big) \tag{D.114}$$

which proves the claim. □

D.9 Matheron's Rule for Gaussian Random Elements

Matheron's rule [Journel and Huijbregts, 1976, Wilson et al., 2020] is a simple trick to sample a conditional Gaussian. Let $(U, V) \sim \mathcal{N}(\mu, \Sigma)$ with

$$\mu = \begin{pmatrix} \mu_U \\ \mu_V \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{UU} & \Sigma_{UV} \\ \Sigma_{VU} & \Sigma_{VV} \end{pmatrix} \tag{D.115}$$

Then $U|V = v \sim \widetilde{U} + \Sigma_{UV}\Sigma_{VV}^{-1}(v - \widetilde{V})$ where $(\widetilde{V}, \widetilde{U}) \sim \mathcal{N}(\mu, \Sigma)$ is independent from (U, V). In other words: We can transform a sample from the joint distribution of (U, V) into a sample of the conditional distribution $U|V = v$.

We want to use Matheron's rule to generate samples from $F(X^*)|F^{1:M} = u$ where $X^* \in \mathcal{X}^{N_*}$ is a set of input locations. Since $F \sim \mathcal{N}(0, C)$ is a GRE, we know that $\big(F(X^*), F^{1:M}\big)$ is jointly Gaussian with mean zero and covariance matrix

$$R := \begin{pmatrix} \mathbb{C}\big[F(X^*), F(X^*)\big] & \mathbb{C}\big[F(X^*), F^{1:M}\big] \\ \mathbb{C}\big[F^{1:M}, F(X^*)\big] & \mathbb{C}\big[F^{1:M}, F^{1:M}\big] \end{pmatrix}. \tag{D.116}$$

We calculate the relevant covariance matrices as

$$\mathbb{C}\big[F(X^*), F(X^*)\big] = r(X^*, X^*) \qquad \mathbb{C}\big[F^{1:M}, F^{1:M}\big] = \Lambda \tag{D.117}$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_M)$ and r is the kernel defined in (6.10). Further, we have

$$\mathbb{C}\big[F^m, F(x_n^*)\big] = \langle Ce_m, k_{x_n^*}(\cdot)\rangle = \lambda_m \langle e_m, k_{x_n^*}(\cdot)\rangle = \lambda_m e_m(x_n^*) \tag{D.118}$$

for all $m = 1, \ldots, M$, $n = 1, \ldots, N^*$ and consequently

$$\mathbb{C}\big[F^{1:M}, F(X^*)\big] = \Lambda e^{1:M}(X^*) \in \mathbb{R}^{M \times N_*}, \tag{D.119}$$

where $(e^{1:M}(X^*))_{m,n} := e_m(x_n^*)$. Matheron's rule therefore takes the form

$$F(X^*)|F^{1:M} = v \sim U + e^{1:M}(X^*)^T(v - V) \tag{D.120}$$

for all $u \in \mathbb{R}^M$ with $(U, V) \sim \mathcal{N}(0, R)$. Naturally, we do not have access to $e^{1:M}(X^*)$, Λ and R. We

therefore use the corresponding approximations

$$\widehat{R} := \begin{pmatrix} \widehat{r}(X^*, X^*) & \widehat{e}^{1:M}(X^*)^T \widehat{\Lambda} \\ \widehat{\Lambda} \widehat{e}^{1:M}(X^*) & \widehat{\Lambda} \end{pmatrix}, \tag{D.121}$$

where $\widehat{\Lambda}$ and $(\widehat{e}^{1:M}(X^*))_{m,n} := \widehat{e}_m(x_n^*)$ $(m = 1, \dots, M, n = 1, \dots, N_*)$ are obtained from the Nyström approximation based on the samples $Z := z_{1:M}$ (cf. Appendix D.7). The kernel matrix is approximated via

$$\widehat{r}(X^*, X^*) := \frac{1}{N_* + M} k(X^*, \widehat{X}) k(\widehat{X}, X^*) \tag{D.122}$$

with $\widehat{X} := (X^*, Z) \in \mathcal{X}^{N_* + M}$. Notice that inclusion of X^* in $\widehat{X}$ leads to $\widehat{r}(X^*, X^*)$ having full rank N^*.

D.10 Asymptotic Analysis of Projected Langevin Sampling

Let $F \sim \mathcal{N}(0, C)$ be a GRE and assume that $-\log p(y|f) = \ell_N(f(x_{1:N}))$ for a function $\ell_N : \mathbb{R}^N \to \mathbb{R}$. Let further $\left(\widehat{F}^{1:M}(t)\right)_{t \geq 0}$ be the solution to the SDE whose components are given as

$$d\widehat{F}^m(t) = -\left(\sum_{n=1}^{N} (\partial_n \ell_N)\Big(\Pr[F(t)](x_{1:N})\Big) e_m(x_n) + \frac{\widehat{F}^m(t)}{\lambda_m} \right) dt + \sqrt{2} dB^m(t), \tag{D.123}$$

where $\Pr[F(t)] = \sum_{m=1}^{M} \widehat{F}^m(t) e_m$. Notice, that for $\ell_N\big(f(x_{1:N})\big) = \sum_{n=1}^{N} c\big(y_n, f(x_n)\big)$ we recover (6.14). We want to show that

$$\widehat{F}^{1:M}(t) \xrightarrow{\mathcal{D}} \tau_\infty \tag{D.124}$$

for $t \to \infty$ where τ_∞ has the potential

$$V_\infty(u) = \ell_N\big(\mu_u(x_{1:N})\big) + \frac{1}{2} u^T {\Lambda_M}^{-1} u + \text{ const.} \tag{D.125}$$

for $u \in \mathbb{R}^M$ where $\mu_u(x_{1:N}) = u^T e^{1:M}(x_{1:N})$ and $\Lambda_M = \operatorname{diag}(\lambda_1, \dots, \lambda_M)$.

Proof. Let V_∞ be the potential above and note that, by the chain-rule, we have

$$\partial_m V_\infty(u) = \sum_{n=1}^{N} \partial_n \ell_N\big(\mu_u(x_{1:N})\big) e_m(x_n) + \frac{u_m}{\widehat{\lambda}_m} \tag{D.126}$$

where ∂_m denotes the derivative with respect to m-th coordinate. Let further $\nabla V_\infty = (\partial_1 V_\infty, \dots, \partial_M V_\infty)^T$ be the gradient of V_∞. We now immediately recognise (D.123) as Langevin diffusion for the potential V, since it holds that

$$d\widehat{F}^m(t) = -\partial_m V\big(\widehat{F}^{1:M}(t)\big) + \sqrt{2} dB^m(t). \tag{D.127}$$

It is well established [Roberts and Tweedie, 1996] that–under mild assumption on V–the limiting distribution of (D.127) is given as τ_∞. □

D.11 Optimal Variational Approximation

We prove a slightly more general result. Indeed, we can derive the optimal variational distribution for arbitrary inducing features U with $U := \big(\langle F, h_m\rangle\big)_{m=1}^M \in \mathbb{R}^M$ for set of functions $\{h_1, \dots, h_M\} \subset H_k$. Theorem 6.6.1 follows for the choice $h_m = e_m$, $m = 1, \dots, M$. Define now for $\tau \in \mathcal{P}_2(\mathbb{R}^m)$ the measure

$$Q_\tau(A) := \int \mathbb{P}\big(F \in A \,|\, U = u\big)\, d\tau(u), \tag{D.128}$$

for $A \in \mathcal{B}(H_k)$ and the set $\mathcal{Q}_M := \{Q_\tau : \tau \in \mathcal{P}_2(\mathbb{R}^M)\} \subset \mathcal{P}_2(H_k)$. We can find a closed form expression for the best posterior approximation in $\mathcal{Q}_M$.

Theorem D.11.1. *Define* Π_M^* *as*

$$\Pi_M^* := \operatorname*{arg\,min}_{Q \in \mathcal{Q}_M} \mathrm{KL}(Q, \Pi^*) \tag{D.129}$$

where Π^* *is the Bayesian posterior. Then* Π_M^* *satisfies* (D.128) *with* $\tau^* \propto \exp(-V^*(u))$ *with*

$$V^*(u) := \mathbb{E}_\xi\Big[\ell_N\big(\mu_u(x_{1:N}) + \sqrt{\Sigma(x_{1:N})}\xi\big)\Big] + \frac{1}{2}\Big(\langle Ch, h\rangle\Big)^{-1} \tag{D.130}$$

for all $u \in \mathbb{R}^M$. *Here,* $\xi \sim \mathcal{N}(0, I_M)$, *and we define*

$$\mu_u(x_{1:N}) := \mathbb{C}\big[F(x_{1:N}), U\big]\mathbb{C}[U, U]^{-1} u \tag{D.131}$$

$$\Sigma(x_{1:N}) := \mathbb{C}\big[F(x_{1:N}), F(x_{1:N})\big] - \mathbb{C}\big[F(x_{1:N}), U\big]\mathbb{C}[U, U]^{-1}\mathbb{C}\big[U, F(x_{1:N})\big] \tag{D.132}$$

$$\langle Ch, h\rangle := \big(\langle Ch_m, h_{m'}\rangle\big)_{m,m'=1}^M \tag{D.133}$$

for all $u \in \mathbb{R}^M$. *Notice that by standard properties of the GRE these covariance terms can be expressed in terms of the covariance operator* C.

Proof. Let $Q \in \mathcal{Q}_M$ be of the form (D.128) with probability measure τ. By Theorem 4 in Wild and Wynne [2022], we know that

$$\mathrm{KL}(Q, \Pi^*) = \mathbb{E}_Q\Big[\ell_N(f(x_{1:N}))\Big] + \mathrm{KL}(\tau, \Pi_U) + \log p(y) \tag{D.134}$$

where Π_U is the prior law of U given as $\mathcal{N}(0, \langle Ch, h\rangle)$. Furthermore, we know that $F(X)|U$ is Gaussian for every $Q \in \mathcal{Q}_M$ by definition $\mathcal{Q}_M$. We therefore know by standard properties of Gaussians that

$$F(x_{1:N})|U = u \sim \mu_u(x_{1:N}) + \sqrt{\Sigma(x_{1:N})}\xi \tag{D.135}$$

for fixed $u \in \mathbb{R}^M$ and a $\xi \sim \mathcal{N}(0, I_M)$. We can therefore condition on $U = u$ and use the tower property of expected values to obtain

$$\mathrm{KL}(Q, \Pi^*) = \int_{\mathbb{R}^M} \mathbb{E}_\xi\Big[\ell_N\big(\mu_u(x_{1:N}) + \sqrt{\Sigma(x_{1:N})}\xi\big)\Big]\, d\tau(u) + \mathrm{KL}(\tau, \Pi_U) + \log p(y). \tag{D.136}$$

We now define $L(\tau) := \int \phi(u)\, d\tau + \mathrm{KL}(\tau, \Pi_U)$ with $\phi(u) := \mathbb{E}_\xi\Big[\ell_N\big(\mu_u(x_{1:N}) + \sqrt{\Sigma(x_{1:N})}\xi\big)\Big]$, $u \in \mathbb{R}^M$ and $\tau \in \mathcal{P}_2(\mathbb{R}^M)$. From the above calculations, it then immediately follows that

$$\min_{Q \in \mathcal{Q}} \mathrm{KL}(Q, \Pi^*) = \min_{\tau} L(\tau) + \log p(y). \tag{D.137}$$

However, the global minimiser of L is well-known and given as $\tau^* \propto \exp(-V^*(u))$ [Knoblauch et al., 2019, Theorem 1] with potential

$$V^*(u) = \phi(u) - \log \frac{d\Pi_U}{du}(u) = \phi(u) + \frac{1}{2}\Big(\langle Ch, h\rangle\Big)^{-1}. \tag{D.138}$$

This proofs our claim. □

D.12 Optimality of Projected Langevin Sampling

Let $\{\lambda_n, e_n\}_{n=1}^\infty$ be the eigenvalue-eigenfunction pairs of the covariance operator C with $\lambda_1 > \lambda_2 > \ldots$ and define $F^m := \langle F, e_m\rangle$, $m = 1, \ldots, M$ and $F^{1:M} := \big(F^1, \ldots, F^M\big)^T$. We simulate according to the PLS algorithm and obtain τ_∞ with potential V_∞ as derived in Appendix D.10. The PLS approximation to the posterior is defined as

$$\widehat{\Pi}(A) := \int \mathbb{P}(F \in A \mid F^{1:M} = u)\, d\tau_\infty(u) \tag{D.139}$$

for $A \in \mathcal{B}(H_k)$. By construction, clearly $\widehat{\Pi} \in \mathcal{Q}_M$ and so it is natural to compare $\widehat{\Pi}$ to the optimal measure Π_M^* of Theorem D.11.1.

Theorem D.12.1. *Assume that $-\log p(y_{1:N}|f) = \ell_N(f(x_{1:N}))$ for a $\kappa-$Lipschitz continuous ($\kappa > 0$) and convex function $\ell_N : \mathbb{R}^N \to \mathbb{R}$. Then, for fixed $x_1, \ldots, x_N \in \mathcal{X}$, we have*

$$\mathrm{KL}\left(\widehat{\Pi}, \Pi_M^*\right) \leq \frac{\kappa^2}{2} tr[\Sigma(x_{1:N})] = \frac{\kappa^2}{2} \sum_{m=M+1}^{\infty} \lambda_m \sum_{n=1}^{N} (e_m(x_n))^2 \tag{D.140}$$

Further, if $X_1, \ldots, X_N$ independently and identically distributed with ν, then

$$\mathbb{E}_{X_{1:N}}\left[\mathrm{KL}\left(\widehat{\Pi}, \Pi_M^*\right)\right] \leq \frac{N\kappa^2}{2} \sum_{m=M+1}^{\infty} \lambda_m^2. \tag{D.141}$$

Proof. Let Π the Gaussian prior measure. Define the map $\Phi : H_k \to \mathbb{R}^M$ as $\Phi(f) := \langle f, e^{1:M} \rangle := \big(\langle f, e_m \rangle\big)_{m=1}^M$. Then by construction we have

$$\Phi \# \widehat{\Pi} = \tau_\infty \qquad \text{and} \qquad \Phi \# \Pi_M^* = \tau^*. \tag{D.142}$$

Furthermore $\widehat{\Pi} \in \mathcal{Q}$ means that $\widehat{\Pi}$ is dominated by the prior measure Π and of the form [Matthews et al., 2016, Wild et al., 2023]

$$\frac{d\widehat{\Pi}}{d\Pi}(f) = \frac{d(\Phi \# \widehat{\Pi})}{d(\Phi \# \Pi)}\big(\phi(f)\big) = \frac{d\tau_\infty}{d(\mathcal{N}(0, \Lambda_M))}\big(\phi(f)\big) \tag{D.143}$$

and similarly for Π_M^*

$$\frac{d\Pi_M^*}{d\Pi}(f) = \frac{d(\Phi \# \Pi_M^*)}{d(\Phi \# \Pi)}\big(\phi(f)\big) = \frac{d\tau^*}{d(\mathcal{N}(0, \Lambda_M))}\big(\phi(f)\big) \tag{D.144}$$

for all $f \in H_k$. Here $\Lambda_M := \mathrm{diag}\big(\lambda_1, \ldots, \lambda_M\big)$, Consequently, we have by standard rules for Radon-Nikodym derivatives that

$$\frac{d\widehat{\Pi}}{d\Pi_M^*}(f) = \frac{d\tau_\infty}{d\tau^*}(\phi(f)) \tag{D.145}$$

for all $f \in H_k$. We can now calculate the KL-divergence as

$$\mathrm{KL}(\widehat{\Pi}, \Pi_M^*) = \int_H \log\left(\frac{d\widehat{\Pi}}{d\Pi_M^*}\right)(f)\, d\widehat{\Pi}(f) \tag{D.146}$$

$$= \int_H \log\left(\frac{d\tau_\infty}{d\tau^*}\right)(\phi(f))\, d\widehat{\Pi}(f) \tag{D.147}$$

$$= \int_{\mathbb{R}^M} \log \frac{d\tau_\infty}{d\tau^*}(u)\, d\tau_\infty(u) \tag{D.148}$$

$$= \mathrm{KL}(\tau_\infty, \tau^*) \tag{D.149}$$

where we applied the change of measure formula. It remains to find an upper bound for the KL-divergence between the two probability measures $\tau_\infty, \tau^* \in \mathcal{P}_2(\mathbb{R}^M)$.

By Theorem D.11.1, we know that τ^* has potential given as

$$V^*(u) = \mathbb{E}_\xi\Big[\ell_N\big(\mu_u(x_{1:N}) + \sqrt{\Sigma(x_{1:N})}\xi\big)\Big] + \frac{1}{2}\Big(\langle Ch, h\rangle\Big)^{-1} \tag{D.150}$$

where $\xi \sim \mathcal{N}(0, I_M)$. For $h_m = e_m$ the mean vector and covariance matrix can be calculated as

$$\mu_u(x_{1:N}) := \mathbb{C}\big[F(x_{1:N}), U\big]\mathbb{C}[U, U]^{-1}u \tag{D.151}$$

$$= \big\langle Ck_{x_{1:N}}(\cdot), e^{1:M}\big\rangle\big(\langle Ce^{1:M}, e^{1:M}\rangle\big)^{-1}u \tag{D.152}$$

$$= e^{1:M}(x_{1:N})^T\Lambda_M(\Lambda_M)^{-1}u \tag{D.153}$$

$$= e^{1:M}(x_{1:N})^T u \tag{D.154}$$

and further

$$\Sigma(x_{1:N}) := \mathbb{C}\big[F(x_{1:N}), F(x_{1:N})\big] - \mathbb{C}\big[F(x_{1:N}), U\big]\mathbb{C}[U, U]^{-1}\mathbb{C}\big[U, F(x_{1:N})\big] \tag{D.155}$$

$$= r(x_{1:N}, x_{1:N}) - e^{1:M}(x_{1:N})^T\Lambda_M e^{1:M}(x_{1:N}) \tag{D.156}$$

where $e^{1:M}(x_{1:N}) := \big(e^m(x_n)\big)_{m,n=1}^N \in \mathbb{R}^{M\times N}$.

Further, the potential of τ_∞ is given as (cf. Appendix D.10)

$$V_\infty(u) = \ell_N(\mu_u(x_{1:N})) + \frac{1}{2}u^T\Lambda_M^{-1}u \tag{D.157}$$

where the second equality follows from the definition of ℓ_N and the calculations for $\mu_u(X)$. Due to the convexity of ℓ_N, we obtain from Jensen's inequality

$$V^*(u) = \mathbb{E}_\xi\Big[\ell_N\big(\mu_u(x_{1:N}) + \sqrt{\Sigma(x_{1:N})}\xi\big)\Big] + \frac{1}{2}u^T\Lambda_M^{-1}u \tag{D.158}$$

$$\geq \ell_N\Big(\mathbb{E}_\xi\big[\mu_u(x_{1:N}) + \sqrt{\Sigma(x_{1:N})}\xi\big]\Big) + \frac{1}{2}u^T\Lambda_M^{-1}u \tag{D.159}$$

$$= V_\infty(u). \tag{D.160}$$

By Lemma 3 in Dalalyan [2017] this implies that

$$\mathrm{KL}(\tau_\infty, \tau^*) \leq \frac{1}{2}\mathbb{E}_U\left[|V^*(U) - V_\infty(U)|^2\right] \tag{D.161}$$

where $U \sim \tau_\infty$. We now calculate

$$\begin{aligned}
\mathbb{E}_U\left[|V^*(U) - V_\infty(U)|^2\right] &= \mathbb{E}_U\left[|\mathbb{E}_\xi\left[\ell_N(\mu_U(x_{1:N})) - \ell_N(\mu_U(x_{1:N}) + \sqrt{\Sigma}\xi)\right]|^2\right] && \text{(D.162)}\\
&\leq \mathbb{E}_U\mathbb{E}_\xi\left[|\ell_N(\mu_U(x_{1:N})) - \ell_N(\mu_U(x_{1:N}) + \sqrt{\Sigma}\xi)|^2\right] && \text{(D.163)}\\
&\leq \kappa^2\mathbb{E}_U\mathbb{E}_\xi\left[||\sqrt{\Sigma}(x_{1:N})\xi)||^2\right] && \text{(D.164)}\\
&= \kappa^2\mathbb{E}\left[||\sqrt{\Sigma}(x_{1:N})\xi)||^2\right] && \text{(D.165)}\\
&= \kappa^2\mathrm{tr}[\Sigma(x_{1:N})], && \text{(D.166)}
\end{aligned}$$

where the first inequality is due to Jensen's inequality and the second inequality uses the Lipschitz continuity of ℓ_N.

We combine the results to obtain

$$\mathrm{KL}\left(\widehat{\Pi}, \Pi_M^*\right) = \mathrm{KL}(\tau_\infty, \tau^*) \leq \frac{1}{2}\kappa^2\mathrm{tr}[\Sigma(x_{1:N})]. \tag{D.167}$$

Furthermore, the trace can be simplified to

$$\begin{aligned}
\mathrm{tr}[\Sigma(x_{1:N})] &= \sum_{n=1}^N r(x_n, x_n) - \sum_{n=1}^M \lambda_m \mathrm{tr}[e_m(x_{1:N})^T e_m(x_{1:N})] && \text{(D.168)}\\
&= \sum_{n=1}^N r(x_n, x_n) - \sum_{m=1}^M \lambda_m \sum_{n=1}^N (e_m(x_n))^2. && \text{(D.169)}
\end{aligned}$$

Furthermore note that

$$\begin{aligned}
r(x, x') &= \langle Ck_x(\cdot), k_{x'}(\cdot)\rangle && \text{(D.170)}\\
&= \sum_{m=1}^\infty \lambda_m \langle k_x, e_m\rangle\langle k_{x'}, e_m\rangle && \text{(D.171)}\\
&= \sum_{m=1}^\infty \lambda_m e_m(x) e_m(x') && \text{(D.172)}
\end{aligned}$$

for all $x, x' \in \mathcal{X}$. We plug this expression back into (D.169) and obtain

$$\mathrm{tr}[\Sigma(x_{1:N})] = \sum_{m=M+1}^\infty \lambda_m \sum_{n=1}^N (e_m(x_n))^2. \tag{D.173}$$

Furthermore we know that for $X_n \sim \nu$ it holds that

$$\mathbb{E}\big[(e_m(X_n))^2\big] = \int (e_m(s))^2 \, d\nu(s) \tag{D.174}$$

$$= \lambda_m, \tag{D.175}$$

due to the isometric isomorphism S introduced in Appendix D.4. See also Steinwart and Christmann [2008, Theorem 4.51]. We combine all calculations and obtain for fixed $x_1, \ldots, x_N$ the inequality

$$\mathrm{KL}\left(\widehat{\Pi}, \Pi_M^*\right) \leq \frac{\kappa^2}{2} \sum_{m=M+1}^{\infty} \lambda_m \sum_{n=1}^{N} (e_m(x_n))^2 \tag{D.176}$$

and further for $X_1, \ldots, X_N \sim \nu$

$$\mathbb{E}_{X_{1:N}}\Big[\mathrm{KL}\left(\widehat{\Pi}, \Pi_M^*\right)\Big] \leq \frac{N\kappa^2}{2} \sum_{m=M+1}^{\infty} \lambda_m^2 \tag{D.177}$$

which proves the claim. □

The assumptions that ℓ_N is Lipschitz continuous and convex is indeed often satisfies in practice. For example, the binary classification loss with logistic link function introduced in Section 6.4.1 is both convex and Lipschitz continuous [Steinwart and Christmann, 2008]. Unfortunately, the squared loss does not satisfy Lipschitz continuity (even though it is convex), but in this case the PLS measure $\widehat{\Pi}$ actually coincides with $\Pi_{|\mathcal{Q}}^*$.

Lemma D.12.2. *Assume that* $\ell_N\big(f(x_{1:N})\big) = \frac{1}{2\sigma^2}||y_{1:N} - f(x_{1:N})||^2$. *Then* $\tau_\infty = \tau^*$ *and consequently* $\widehat{\Pi} = \Pi_M^*$.

Proof. According to Theorem D.11.1 the potential of τ^* is given as ($h_m = e_m$)

$$V^*(u) = \mathbb{E}_\xi\Big[\ell_N\big(\mu_u(x_{1:N}) + \sqrt{\Sigma(x_{1:N})}\xi\big)\Big] + \frac{1}{2}u^T\Big(\langle Ch, h\rangle\Big)^{-1}u \tag{D.178}$$

$$= \frac{1}{2\sigma^2}\mathbb{E}_\xi\left[||Y_{1:N} - \mu_u(x_{1:N}) + \sqrt{\Sigma(x_{1:N})}\xi||^2\right] + \frac{1}{2}u^T\Lambda_M^{-1}u \tag{D.179}$$

where $\xi \sim \mathcal{N}(0, I_M)$ and $\Lambda_M = \mathrm{diag}(\lambda_1, \ldots, \lambda_M)$. Note that

$$Y_{1:N} - \mu_u(x_{1:N}) + \sqrt{\Sigma(x_{1:N})}\xi \sim \mathcal{N}\big(y_{1:N} - \mu_u(x_{1:N}), \Sigma(x_{1:N})\big) \tag{D.180}$$

and so by Equation 378 in Petersen et al. [2008] we obtain

$$V^*(u) = \text{tr}[\Sigma(x_{1:N})] + \frac{1}{2\sigma^2}||Y_{1:N} - \mu_u(x_{1:N})||^2 + u^T\Lambda_M^{-1}u \tag{D.181}$$

$$= \frac{1}{2\sigma^2}||Y_{1:N} - \mu_u(x_{1:N})||^2 + u^T\Lambda_M^{-1}u + \text{ const.} \tag{D.182}$$

$$= \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mu_u(x_n))^2 + u^T\Lambda_m^{-1}u + \text{ const.} \tag{D.183}$$

$$= V_\infty(u) + \text{ const.}, \tag{D.184}$$

where the last equality follows from $\mu_u(x_{1:N}) = e^{1:M}(x_{1:N})^T u$ which we calculated in (D.154). We therefore conclude that V^* and V_∞ are the same up to an additive constant in dependent of u which implies that $\tau^* = \tau_\infty$. Since both $\widehat{\Pi}$ and Π_M^* are in $\mathcal{Q}_M$ we immediately conclude that that $\widehat{\Pi} = \Pi_M^*$. □

D.13 Projected Langevin Sampling for Other Inducing Functions

The idea of PLS can be generalised to inducing functions other than $\widehat{e}^{1:M}$. To this end, let now $P : H_k \to H_M$ be the orthogonal projection onto $H_M := \text{span}\big(\{h_1, \ldots, h_M\}\big) \subset H_k$, $h = (h_1, \ldots, h_M)$ given as

$$Pf = h(\cdot)^T\langle h, h\rangle^{-1}\langle f, h\rangle., \tag{D.185}$$

where $(\langle h, h\rangle)_{m,m'} = \langle h_m, h_{m'}\rangle$ and $(\langle f, h\rangle)_m = \langle f, h_m\rangle$ for all $m = 1, \ldots, M$. We first prove that P is indeed given by expression (D.185).

Lemma D.13.1. *The orthogonal operator $P : H_k \to H_M$ is given by expression (D.185).*

Proof. By definition of the orthogonal projection P onto H_M satisfies

$$||f - P(f)||_k = \arg\min_{g \in H_M} ||f - g||_k, \tag{D.186}$$

where $||\cdot||_k$ is the norm induced by the RKHS inner product. Every element in $g \in H_M$ can, by definition, be written as

$$g(x) = \sum_{m=1}^{M} \alpha_m h_m(x) \tag{D.187}$$

for $\alpha = (\alpha_1, \ldots, \alpha_M) \in \mathbb{R}^M$. Exploiting this fact leads to an finite-dimensional optimisation problem

for $\alpha \in \mathbb{R}^M$ with solution

$$\alpha^* = \langle h, h \rangle^{-1} \langle f, h \rangle. \tag{D.188}$$

and hence $P(f) = h(\cdot)^T \alpha^*$. □

The next step is to derive an evolution equation for the inducing features $U := (U_1, \dots, U_M)^T$ with $U_m := \langle F, h_m \rangle$ $(m = 1, \dots, M)$ from the projected Langevin SDE which is given as

$$dF(t) = -\big(D\ell\big(P(F(t))\big) + C^{-1}F(t)\big)\, dt + \sqrt{2} dW(t). \tag{D.189}$$

Theorem D.13.2. *Define $U(t) := \langle F(t), h_m \rangle$ then $U(t)$ satisfies the SDE in $\mathbb{R}^M$ given as*

$$dU(t) = -\, h(X)\big(\partial_2 c\big)\big(Y, P[F_t](X)\big)dt - \langle C^{-1}F(t), h\rangle dt + \sqrt{2\langle h, h\rangle} dB(t).$$

Here, we denote $h(X) := \big(h_m(x_n)\big)_{m,n=1}^{M,N} \in \mathbb{R}^{M\times N}$, $\big(\partial_2(Y, P[F_t](X))\big)_n := \partial_2 c(y_n, (P[F_t](X))_n)$, $P[F_t](X) = h(X)^T \langle h, h\rangle^{-1} U(t)$, $(\langle C^{-1}F(t), h\rangle)_m = \langle C^{-1}F(t), h_m\rangle)$ for $m = 1, \dots, M$, $n = 1, \dots, N$. Furthermore $B(t)$ is a standard Brownian motion in $\mathbb{R}^M$ and $\sqrt{\langle h, h\rangle}$ the square-root of the matrix $\langle h, h\rangle$.

Proof. We apply Ito's Rule to (D.189) [Da Prato and Zabczyk, 2014, Chapter 4.4] with $\phi : H \to \mathbb{R}^M$ defined as $\phi(f) := \langle f, h\rangle$ which gives the dynamics

$$\begin{aligned} dU(t) &= d\phi\big(F(t)\big) && \text{(D.190)}\\ &= -\Big(\Big\langle D\ell\big(P[F(t)]\big), h\Big\rangle + \big\langle C^{-1}F(t), h\big\rangle\Big)\, dt + \sqrt{2\langle h, h\rangle} dB(t). && \text{(D.191)} \end{aligned}$$

For a loss ℓ of the form (6.6), we can simplify further by applying the chain-rule for Fréchet derivatives

$$\Big\langle D\ell\big(P[F(t)]\big), h\Big\rangle = h(X)^T \big(\partial_2 c\big)\big(Y, (P[F_t])(X)\big) \tag{D.192}$$

which concludes the proof. □

The SDE for $U(t)$ in Theorem (D.13.2) can not be simulated, since we do not have access to $\langle C^{-1}F(t), h\rangle$. However, for the specific choice $h_m = k(z_m, \cdot)$ where $Z = (z_1, \dots, z_M)$ are samples from ν, we can

find an approximation for this term.

Lemma D.13.3. *Let $z_1, \dots, z_m$ be iid samples from ν, $h = k_Z(\cdot)$ and let C be the covariance operator defined in (6.9). Then*

$$\langle C^{-1} f, k_Z \rangle \approx M k(Z, Z)^{-1} f(Z) \tag{D.193}$$

for all $f \in H_k$.

Proof. Let now $h_m = k(z_m, \cdot)$ and hence $h = k_Z(\cdot)$. By reproducing property we have

$$\langle C^{-1} f, h \rangle = C^{-1} f(Z). \tag{D.194}$$

We now use a standard Monte Carlo approach by assuming that

$$\frac{1}{M} \sum_{m=1}^{M} \delta_{z_m} \approx \nu. \tag{D.195}$$

Consequently, we approximate $C : H \to H$ via for all $x \in \mathcal{X}$ via

$$\begin{aligned} Cf(x) &= \int k(x, x') f(x') \, d\nu(x') && \text{(D.196)} \\ &\approx \frac{1}{M} \sum_{m=1}^{M} k(x, z_m) f(z_m) && \text{(D.197)} \\ &=: \frac{1}{M} k(x, Z) f(Z) && \text{(D.198)} \end{aligned}$$

which leads to $Cf(Z) \approx \frac{1}{M} k(Z, Z) f(Z)$ and ultimately to

$$(C^{-1} f)(Z) \approx M k(Z, Z)^{-1} f(Z). \tag{D.199}$$

This concludes the proof. □

The combination of Theorem D.13.2 and Lemma D.13.3 leads to a (approximate) SDE for $U(t)$ given as

$$\begin{aligned} dU(t) = &- k(Z, X) \big(\partial_2 c\big) \big(Y, k(X, Z) k(Z, Z)^{-1} U(t)\big) dt \\ &- M k(Z, Z)^{-1} U(t) dt + \sqrt{2k(Z, Z)} dB(t) \end{aligned} \tag{D.200}$$

which can be simulated in $\mathbb{R}^M$. For large $t > 0$ we hope that $U(t) \approx U|Y$. Furthermore, by Matheron's Rule, we can transform the samples from U into posterior samples (cf. Appendix D.9). The covariance

matrices are given as

$$\mathbb{C}(\langle F, k_{x^*}\rangle, \langle F, k_{x^*}\rangle) = r(x^*, x^*), \quad \mathbb{C}(\langle F, k_{x^*}\rangle, U) = r(x^*, Z) \qquad \mathbb{C}[U, U] = r(Z, Z), \tag{D.201}$$

since $G(x) := \langle F, k(x, \cdot)\rangle$ is a GP with kernel r (cf. Lemma D.4.2).

Furthermore, we can also obtain the asymptotic distribution of the SDE $U(t)$ in closed form, since it is a preconditioned Langevin equation.

Theorem D.13.4. *Let $\big(U(t)\big)_{t\geq 0}$ be the solution to the SDE in (D.200) with $U(0) = U_0$ for a given initial value $U_0 \in \mathbb{R}^M$. Then $U(t) \xrightarrow{\mathcal{D}} \widehat{Q}_U$ for $t \to \infty$ where $\widehat{Q}_U(du) = \widehat{q}_U(u)du$ with*

$$\widehat{q}_U(u) \propto \exp\left(-\sum_{n=1}^{N} c\big(y_n, k_Z(x_n)^T k(Z,Z)^{-1}u\big) - \frac{M}{2} u^T k(Z,Z)^{-2} u\right), \tag{D.202}$$

for $u \in \mathbb{R}^M$.

Proof. First, define $A := k(Z, Z) \in \mathbb{R}^{M\times M}$. Then the SDE in (D.200) can be rewritten as

$$dU(t) = -A\Big(A^{-1}k_Z(X)\partial_2\big(Y, k_Z(x_n)^T A^{-1}U(t)\big) + MA^{-1}k(Z,Z)^{-1}U(t)\Big)dt + \sqrt{2A}d\beta(t), \tag{D.203}$$

Further, we define the potential $V : \mathbb{R}^M \to \mathbb{R}$ as

$$V(u) = \sum_{n=1}^{N} c(y_n, k_Z(x_n)^T A^{-1}u) + \frac{M}{2} u^T A^{-2} u \tag{D.204}$$

and calculate the gradient ∇V as

$$\nabla V(u) = A^{-1}k(Z, X)(\partial_2 c)(Y, k(X, Z)A^{-1}u) + MA^{-2}u. \tag{D.205}$$

The SDE (D.200) can therefore be rewritten as

$$dU(t) = -A\nabla V\big(U(t)\big)dt + \sqrt{2A}d\beta(t) \tag{D.206}$$

which is the preconditioned Langevin diffusion with potential V. It is known [Bhattacharya and Jiang, 2023] that

$$U(t) \xrightarrow{\mathcal{D}} \frac{1}{\kappa}\exp(-V(u)) \tag{D.207}$$

for $t \to \infty$ where $\kappa := \int \exp(-V(u))\, du$ is the normalising constant. □

D.14 Implementation Details

D.14.1 Hyperparameter Selection

For fair comparison, we shared the same hyperparameters for r across PLS and SVGP. We used an ARD kernel for k when constructing r, tuning the hyperparameters of k. For datasets with $N \leq 2000$ observations (or $N \leq 1000$ for classification), we learned the hyperparameters of k by maximising the exact marginal log-likelihood of a GP with k and a mean zero prior. For a data set with more than 2000 (or 1000 for classification) observations we use the following heuristic taken from Lin et al. [2024]:

1. Randomly select a centroid uniformely at random the training data.
2. Select a subset of size 2000 (or 1000 for classification) with the smallest Euclidean distance to the centroid.
3. Learn kernel hyperparmeters on the data through regression GP marginal likelihood with kernel k and mean zero prior for this subset.

Repeat the above procedure 10 (or 5 for classification) and average the learned hyperparameters

The inducing points $z_1, \dots z_M \in \mathcal{X}$ used for the Nyström method in Appendix D.7 were selected following the greedy variance selection method in Burt et al. [2020b] and Chen et al. [2018].

D.14.2 Projected Langevin Sampling Algorithm

Let $\eta > 0$ be the step-size and $T > 0$ be the the end time of our simulation. Define now $U(t) := \widehat{F}^{1:M}(t)$ where $\widehat{F}^{1:M}(t)$ is the solution to the SDE (D.123) and $t_i := i\eta$ for $i = 0, \dots, I$ with $I = \lfloor T/\eta \rfloor$ and further

$$\widehat{U}(0) \sim Q_0 \tag{D.208}$$

$$\widehat{U}(t_{i+1}) := \widehat{U}(t_i) - \eta \widehat{e}^{1:M}(X)\big(\partial_2 c\big)\big(Y, \widehat{e}^{1:M}(X)^T \widehat{U}(t_i)\big) - \eta \widehat{\Lambda}^{-1} \widehat{U}(t_i) + \sqrt{2\eta}\xi_i, \tag{D.209}$$

where $\eta > 0$ is the step size and $\{\xi_i\}_{i=1}^I$ are i.i.d. $\mathcal{N}(0, I_M)$. Here, we define

$$(\partial_2 c)(Y, \widehat{Y}) := \Big(\partial_2 c(y_n, \widehat{y}_n)\Big)_{n=1}^N \in \mathbb{R}^N, \qquad \widehat{e}^{1:M}(X) := \Big(\widehat{e}^m(x_n)\Big)_{m,n=1}^{M,N} \in \mathbb{R}^{M \times N} \tag{D.210}$$

for all $Y, \widehat{Y} \in \mathbb{R}^N$ and $X = (x_1, \dots, x_N) \in \mathcal{X}^N$.

Notice that (D.209) is the Euler-Maruyama discretisation of the SDE (D.123) and therefore $\widehat{U}(t_i) \approx U(t_i) = \widehat{F}^{1:M}(t_i)$ for small enough η.

D.14.3 Likelihood Functions

Regression As discussed in Section 6.4 the Gaussian likelihood $p(y|f) = \mathcal{N}(f(X), \sigma^2 I_N)$ corresponds to the choice $c : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with $c(y, \widehat{y}) = \frac{1}{2\sigma^2}(y - \widehat{y})^2$ and consequently

$$\partial_2 c(y, \widehat{y}) = \frac{1}{\sigma^2}(\widehat{y} - y). \tag{D.211}$$

Classification In binary classification, we assume a Bernoulli data, i.e.

$$Y_n|F = f \sim \text{Bernoulli}\big(\phi(f(x_n))\big). \tag{D.212}$$

This gives rise to the cost $c : \{0, 1\} \times \mathbb{R} \to \mathbb{R}$ with

$$c(y, \widehat{y}) = -\log p(y|f) = -y \log \phi(\widehat{y}) - (1 - y) \log\big(1 - \phi(\widehat{y})\big), \tag{D.213}$$

where $\widehat{y} = f(x)$. For our experiments we use the logistic function $\phi(\widehat{y}) = \big(1 + \exp(-\widehat{y})\big)^{-1}$. Due to the well-known property $\phi'(\widehat{y}) = \phi(\widehat{y})\big(1 - \phi(\widehat{y})\big)$ we obtain

$$\partial_2 c(y, \widehat{y}) = -y\big(1 - \phi(\widehat{y})\big) + (1 - y)\phi(\widehat{y}) \tag{D.214}$$

for $y \in \{0, 1\}$, $\widehat{y} \in \mathbb{R}$.

Poisson Model We choose a Poisson regression model where we assume

$$Y_n|F = f \sim \text{Poisson}\big((f(x_n)^2\big). \tag{D.215}$$

This gives rise to the cost

$$c(y_n, f(x_n)) = -2y_n \log |f(x_n)| + \big(f(x_n)\big)2 \tag{D.216}$$

$$\partial_2 c(y_n, f(x_n)) = -2y_n \frac{\text{sign}(f(x_n))}{|f(x_n)|} + 2f(x_n) \tag{D.217}$$

$$= -\frac{2y_n}{f(x_n)} + 2f(x_n) \tag{D.218}$$

for $y_n \in \mathbb{N}_0$ and $f(x_n) \in \mathbb{R}$.

Algorithm 1: Projected Langevin Sampling

Input: input data $x_{1:N}$, targets $y_{1:N}$, kernel k, inducing points $z_{1:M}$, step size $\eta > 0$, time horizon T, initialisation prob. measure $Q_0 \in \mathcal{P}(\mathcal{X})$, number of samples J, new points $x^*_{1:N_*}$

Result: Samples $F_1(x^*_{1:N_*}), \dots, F_J(x^*_{1:N_*}) \approx F(x^*_{1:N_*})|y_{1:N}$

for $j = 1, \dots, J$ **do**

Initialise $\widehat{U}_j(t_0) \sim Q_0$

for $i = 0, \dots, T/\eta - 1$ **do**

Generate $\widehat{U}_j(t_{i+1})$ from $\widehat{U}_j(t_i)$ according to the update rule in (D.209)

Sample $\big(G_j(x^*_{1:N_*}), \langle G_j, \widehat{e}^{1:M} \rangle\big) \sim \mathcal{N}(0, R_{N_*,M})$ with

$$R_{N_*,M} := \begin{bmatrix} r(x^*_{1:N_*}, x^*_{1:N_*}) & \widehat{e}^{1:M}(x^*_{1:N_*})^T \widehat{\Lambda}_M \\ \widehat{\Lambda}_M \widehat{e}^{1:M}(x^*_{1:N_*}) & \widehat{\Lambda}_M \end{bmatrix} \in \mathbb{R}^{(N_*+M)\times(N_*+M)} \tag{D.219}$$

Calculate

$$F_j(x^*_{1:N_*}) = G_j(x^*_{1:N_*}) + \widehat{e}^{1:M}(x^*_{1:N_*})^\top \Big(\widehat{U}_j(T) - \langle G_j, \widehat{e}^{1:M} \rangle\Big)$$

Here, $\widehat{e}^{1:M}(x^*_{1:N_*}) \in \mathbb{R}^{M\times N_*}$ is the matrix whose entry at (m, n) is $\widehat{e}_m(x^*_n)$ and $\widehat{\Lambda}_M := \mathrm{diag}(\widehat{\lambda}_1, \dots, \widehat{\lambda}_M) \in \mathbb{R}^{M\times M}$ is the diagonal matrix with entries $\widehat{\lambda}_1, \dots, \widehat{\lambda}_M$ (see Appendix D.7 for a definition).

D.14.4 Time and Space Complexity

This section discusses the time and space complexity requirements of producing $J \in \mathbb{N}$ posterior samples from $F|Y$ with our method.

Training We calculate the spectral decomposition of $\frac{1}{M}k(z_{1:M}, z_{1:M})$ and store the result which can be done in $\mathcal{O}(M^3)$. The update in (D.209) requires only matrix multiplications which are each dominated by $\mathcal{O}(NM)$.

The total costs therefore are $\mathcal{O}\big(M^3 + JNM\big)$. These costs could be reduced further by batch-approximations of the gradient. However, the data set considered in this paper serve illustration purposes only and a batch-approximation was not required.

Prediction Let $x^*_{1:N_*} \in \mathcal{X}^{N_*}$ be a set of input points for which we want to generate posterior samples, i.e. $F_j(x^*_{1:N_*}) \approx F(x^*_{1:N_*})|y_{1:N}$. Generating one sample in (D.219) requires jointly sampling $\big(G_j(x^*_{1:N_*}), \langle G_j, \widehat{e}^{1:M} \rangle\big)$ from a multivariate Gaussian which is $\mathcal{O}\big((N_* + M)^3\big)$ for calculating the Cholesky (or the spectral) decomposition once and then $\mathcal{O}\big((N_* + M)^2\big)$ for generating samples. The matrix multiplication is $\mathcal{O}(N_*M)$. Hence we pay $\mathcal{O}\big((N_* + M)^3\big)$ upfront for the Cholesky decomposition and then $\mathcal{O}\big(J(N_* + M)^2\big)$.

For special kernels, this could be further improved by using the exploiting the ideas discussed in Wilson

et al. [2020].

Space Complexity Space complexity in our implementation is $\mathcal{O}(NM+M^2)$ to store the $k(z_{1:M}, x_{1:N})$ and $k(z_{1:M}, z_{1:M})$ matrix.