

# The Ridgelet Prior: A Covariance Function Approach to Prior Specification for Bayesian Neural Networks

**Takuo Matsubara**

*School of Mathematics, Statistics & Physics  
Newcastle University*

*Newcastle upon Tyne, NE1 7RU, United Kingdom*

TMATSUBARA@TURING.AC.UK

**Chris J. Oates**

*School of Mathematics, Statistics & Physics  
Newcastle University*

*Newcastle upon Tyne, NE1 7RU, United Kingdom*

CHRIS.OATES@NCL.AC.UK

**François-Xavier Briol**

*Department of Statistical Science*

*University College London*

*London, WC1E 6BT, United Kingdom*

F.BRIOL@UCL.AC.UK

**Editor:** Pierre Alquier

## Abstract

Bayesian neural networks attempt to combine the strong predictive performance of neural networks with formal quantification of uncertainty associated with the predictive output in the Bayesian framework. However, it remains unclear how to endow the parameters of the network with a prior distribution that is meaningful when lifted into the output space of the network. A possible solution is proposed that enables the user to posit an appropriate Gaussian process covariance function for the task at hand. Our approach constructs a prior distribution for the parameters of the network, called a *ridgelet prior*, that approximates the posited Gaussian process in the output space of the network. In contrast to existing work on the connection between neural networks and Gaussian processes, our analysis is non-asymptotic, with finite sample-size error bounds provided. This establishes the universality property that a Bayesian neural network can approximate any Gaussian process whose covariance function is sufficiently regular. Our experimental assessment is limited to a proof-of-concept, where we demonstrate that the ridgelet prior can out-perform an unstructured prior on regression problems for which a suitable Gaussian process prior can be provided.

**Keywords:** Bayesian neural networks, Gaussian processes, prior selection, ridgelet transform, statistical learning theory

## 1. Introduction

Neural networks are beginning to be adopted in a range of sensitive application areas such as healthcare (Topol, 2019), social care (Serrano and Bajo, 2019), and the justice system (Tortora et al., 2020), where the accuracy and reliability of their predictive output demands careful assessment. This problem lends itself naturally to the Bayesian paradigm and there has been a resurgence in interest in Bayesian neural networks (BNNs), originally introduced

and studied in (Buntine and Weigend, 1991; Mackay, 1992; Neal, 1995). BNNs use the language of probability to express uncertainty regarding the “true” value of the parameters in the neural network, initially by assigning a *prior* distribution over the space of possible parameter configurations and then updating this distribution on the basis of a training dataset. The resulting *posterior* distribution over the parameter space implies an associated predictive distribution for the output of the neural network, assigning probabilities to each of the possible values that could be taken by the output of the network. This predictive distribution carries the formal semantics of the Bayesian framework and can be used to describe epistemic uncertainty associated with the phenomena being modelled.

Attached to any probabilistic quantification of uncertainty are *semantics*, which describe how probabilities should be interpreted (e.g. are these probabilities epistemic or aleatoric; whose belief is being quantified; what assumptions are premised?). As for any Bayesian model, the semantics of the posterior predictive distribution are largely inherited from the semantics of the prior distribution, which is typically a representation of a user’s subjective belief about the unknown “true” values of parameters in the model. This represents a challenge for BNNs, as a user cannot easily specify their prior belief at the level of the parameters of the network in general settings where the influence of each parameter on the network’s output can be difficult to understand. Furthermore, the total number of parameters can go from a few dozens to several million or more, rendering careful selection of priors for each parameter impractical. This has led some researchers to propose *ad hoc* choices for the prior distribution, which will be reviewed in Section 2 (see also Nalisnick, 2018). Such *ad hoc* choices of prior appear to severely limit interpretability of the semantics of the BNN. It has also been reported that such priors can have negative consequences for the predictive performance of BNNs (Yao et al., 2019).

The development of interpretable prior distributions for BNNs is an active area of research that, if adequately solved, has the potential to substantially advance methodology for neural networks. Potential benefits include:

- **Fewer Data Required:** BNNs are “data hungry” models; their large number of parameters means that a large number of data are required for the posterior to concentrate on a suitable configuration of parameter values. The inclusion of domain knowledge in the prior distribution could be helpful in reducing the effective degrees of freedom in the parameter space, mitigating the requirement for a large training dataset.
- **Faster Computation:** The use of an *ad hoc* prior distribution can lead to a posterior distribution that is highly multi-modal (Pourzanjani et al., 2017), creating challenges for computation (e.g. using variational inference or Markov chain Monte Carlo; MCMC). The inclusion of domain knowledge could be expected to counteract (to some extent) the multi-modality issue by breaking some of the symmetries present in the parametrisation of the network.
- **Lower Generalisation Error:** An important issue with BNNs is that their out-of-sample performance can be poor when an *ad hoc* prior is used. These issues have led several authors to question the usefulness of the BNNs; see Mitros and Namee (2019) and Wenzel et al. (2020). Model predictions are strongly driven by the prior and we

therefore expect inclusion of domain knowledge to be an important factor in improving the generalisation performance of BNNs.

In this paper we do *not* claim to provide a solution to the problem of prior selection that enjoys all the benefits just discussed. Such an endeavour would require very extensive (and application-specific) empirical investigation, which is not our focus in this work. Rather, this paper proposes and studies a novel approach to prior specification that operates at the level of the output of the neural network and, in doing so, provides a route for expert knowledge on the phenomenon being modelled to be probabilistically encoded. The construction that we present is stylised to admit a detailed theoretical analysis and therefore the empirical results in this paper are limited to a proof-of-concept. In subsequent work we will discuss generalisations of the construction that may be more amenable to practical applications, for example by reducing the number of hidden units that may be required.

Our analysis can be viewed in the context of a recent line of research which focuses on the predictive distribution as a function of the prior distribution on network parameters (Flam-Shepherd et al., 2017; Hafner et al., 2020; Pearce et al., 2020; Sun et al., 2019). These papers propose to reduce the problem of prior selection for BNNs to the somewhat easier problem of prior selection for Gaussian processes (GPs). The approach studied in these papers, and also adopted in the present paper, can be summarised as follows: (i) Elicit a GP model that encodes domain knowledge for the problem at hand, (ii) Select a prior for the parameters of the BNN such that the output of the BNN in some sense “closely approximates” the GP. This high-level approach is appealing since it provides a direct connection between the established literature on covariance modelling for GPs (Duvenaud, 2014; Rasmussen and Williams, 2006; Stein, 1999) and the literature on uncertainty quantification using a BNN. For instance, existing covariance models can be used to encode *a priori* assumptions of amplitude, smoothness, periodicity and so on as required. Moreover, the number of parameters required to elicit a GP (i.e. the parameters of the mean and covariance functions) is typically much smaller than the number of parameters in a BNN.

Existing work on this topic falls into two categories. In the first, the prior is selected in order to minimise a variational objective between the BNN and the target GP (Flam-Shepherd et al., 2017; Hafner et al., 2020; Sun et al., 2019). Although some of the more recent approaches have demonstrated promising empirical results, all lack theoretical guarantees. In addition, these approaches often constrain the user to use a particular algorithm for posterior approximation (such as variational inference), or require the need to see some of the training data in order to construct the prior model. The second approach consists of carefully adapting the architecture of the BNN to ensure convergence to the GP via a central limit theorem argument (Pearce et al., 2020). This approach is particularly efficient, but requires deriving a new BNN architecture for every GP covariance function and in this sense may be considered impractical.

In this paper we propose the *ridgelet prior*, a novel method to construct interpretable prior distributions for BNNs. It follows the previous two-stage approach, but remedies several issues with existing approaches:

- **Universal Approximation:** The ridgelet prior can be used for a BNN to approximate *any* GP of interest (provided generic regularity conditions are satisfied) in the prior predictive without the need to modify the architecture of the network.

- **Finite Sample-Size Error Bounds:** Approximation error bounds are obtained which are valid for a finite-dimensional parametrisation of the network, as opposed to relying on asymptotic results such as a central limit theorem.
- **Compatibility:** The prior can be used within most existing algorithms for posterior approximation, such as variational inference or MCMC.
- **No Optimisation Required:** The ridgelet prior does not require access to any part of the dataset in its construction and is straight-forward to implement (e.g. it does not require any numerical optimisation routine).

To construct the ridgelet prior, we build on existing analysis of the *ridgelet transform* (Candès, 1998; Murata, 1996; Sonoda and Murata, 2017), which was used to study the consistency of (non-Bayesian) neural networks. In particular, we derive a novel result for numerical approximation using a finite-bandwidth version of the ridgelet transform, presented in Theorem 1, that may be of independent interest. The ridgelet prior is defined for neural networks with  $L \geq 1$  hidden layers but our theoretical analysis focuses on the “shallow” case of  $L = 1$  hidden layer, which is nevertheless a sufficiently rich setting for our consistency results to hold.

The remainder of the paper is structured as follows: Section 2 reviews common prior choices for BNNs and known connections between BNNs and GPs. Section 3 presents the ridgelet prior in full detail. Our theoretical guarantees for the ridgelet prior are outlined in Section 4. A proof-of-concept empirical assessment is contained Section 6. Section 7 summarises the main conclusions of this work. Code to reproduce all results in this paper can be downloaded from: <https://github.com/takuomatsubara/BNN-RidgeletPrior>.

## 2. Background

To begin we briefly introduce notation for GPs and BNNs, discussing the issue of prior specification for these models.

### 2.1 Prior Specification and Covariance Functions

This paper focusses on the problem of approximating a deterministic function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  using a BNN. This problem is fundamental and underpins algorithms for regression and classification. The Bayesian approach is to model  $f$  as a stochastic process (also called “random function”)  $f : \mathbb{R}^d \times \Theta \rightarrow \mathbb{R}$ , where  $\Theta$  is a measurable parameter space on which a *prior* probability distribution is elicited, denoted  $\mathbb{P}$ . The set  $\Theta$  may be either finite or infinite-dimensional. In either case,  $\theta \mapsto f(\cdot, \theta)$  is a random variable taking values in the vector space of real-valued functions on  $\mathbb{R}^d$ . The combination of a dataset of size  $n$  and Bayes’ rule are used to constrain, in a statistical sense, this distribution on  $\Theta$ , to produce a posterior  $\mathbb{P}_n$  that is absolutely continuous with respect to  $\mathbb{P}$ . If the model is well-specified, then there exists an element  $\theta^\dagger \in \Theta$  such that  $f(\cdot, \theta^\dagger) = f(\cdot)$  and, if the Bayesian procedure is consistent,  $\mathbb{P}_n$  will converge (in an appropriate sense) to a point mass on  $\theta^\dagger$  in the  $n \rightarrow \infty$  limit.

In practice, Bayesian inference requires that a suitable prior distribution  $\mathbb{P}$  is elicited. Stochastic processes are intuitively described by their moments, and these can be used

by a domain expert to elicit  $\mathbb{P}$ . The first two moments are given by the mean function  $m : \mathbb{R}^d \rightarrow \mathbb{R}$  and the covariance function  $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ , given pointwise by

$$m(\mathbf{x}) := \int_{\Theta} f(\mathbf{x}, \theta) d\mathbb{P}(\theta), \quad k(\mathbf{x}, \mathbf{x}') := \int_{\Theta} (f(\mathbf{x}, \theta) - m(\mathbf{x}))(f(\mathbf{x}', \theta) - m(\mathbf{x}')) d\mathbb{P}(\theta).$$

for all  $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ . GPs and BNNs are examples of stochastic processes that can be used. In the case of a GP, the first two moments completely characterise  $\mathbb{P}$ . Indeed, under the conditions of Mercer’s theorem (see e.g. Section 4.5 of Steinwart and Christmann, 2008),

$$f(\mathbf{x}, \theta) = m(\mathbf{x}) + \sum_{i=1}^{\dim(\Theta)} \theta_i \varphi_i(\mathbf{x}), \quad \theta_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$$

where the functions  $\varphi_i : \mathbb{R}^d \rightarrow \mathbb{R}$  are obtained from the Mercer decomposition of the covariance function  $k(\mathbf{x}, \mathbf{x}') = \sum_i \varphi_i(\mathbf{x})\varphi_i(\mathbf{x}')$  and  $\dim(\Theta)$  denotes the dimension of  $\Theta$ . The shorthand notation  $\mathcal{GP}(m, k)$  is often used to denote this GP. The *kernel trick* enables explicit computation with the  $\theta_i$  and  $\varphi_i$  to be avoided, so that the user can specify the mean and covariance functions and, in doing so,  $\mathbb{P}$  is implicitly defined. There is a well-established literature on covariance modelling (Duvenaud, 2014; Rasmussen and Williams, 2006; Stein, 1999) for GPs. For BNNs, however, there is no analogue of the kernel trick and it is unclear how to construct a prior  $\mathbb{P}$  for the parameters  $\theta$  of the BNN that is in agreement with moments that have been expert-elicited.

Fix a function  $\phi : \mathbb{R} \rightarrow \mathbb{R}$ , which we will call *activation function*. In this paper a BNN with  $L \geq 1$  *hidden layers* is understood to be a stochastic process with functional form

$$f(\mathbf{x}, \theta) = \sum_{j=1}^{N_L} w_{1,j}^L \phi(z_j^L(\mathbf{x})), \quad z_i^l(\mathbf{x}) := b_i^{l-1} + \sum_{j=1}^{N_{l-1}} w_{i,j}^{l-1} \phi(z_j^{l-1}(\mathbf{x})), \quad l = 2, \dots, L \quad (1)$$

where  $N_l := \dim(\mathbf{z}^l(\mathbf{x}))$  is the number of nodes in the  $l$ th layer and the edge case is the *input layer*

$$z_i^1(\mathbf{x}) := b_i^0 + \sum_{j=1}^d w_{i,j}^0 x_j,$$

The parameters  $\theta$  of the BNN consists of the *weights*  $w_{i,j}^l \in \mathbb{R}$  and the *biases*  $b_i^l \in \mathbb{R}$  of each layer  $l = 0, \dots, L$ , where the  $L$ th layer’s bias is excluded in our definition. Common examples of activation functions include the rectified linear unit (ReLU)  $\phi(x) = \max(0, x)$ , logistic  $\phi(x) = 1/(1 + \exp(-x))$ , hyperbolic tangent  $\phi(x) = \tanh(x)$  and the Gaussian  $\phi(x) = \exp(-x^2)$ . In all cases the complexity of the mapping  $\theta \mapsto f(\cdot, \theta)$  in (1) makes prior specification challenging, since it is difficult to ensure that a distribution on the parameters  $\theta$  will be meaningful when lifted to the output space of the neural network.

## 2.2 The Covariance of Bayesian Neural Networks

Here we discuss existing choices for the prior  $\mathbb{P}$  on  $\theta$  in a BNN, which are motivated by the covariance structure that they induce on the output space of the neural network. This is a rapidly evolving field and a full review requires a paper in itself; we provide a succinct summary and refer the reader to the survey in Nalisnick (2018).

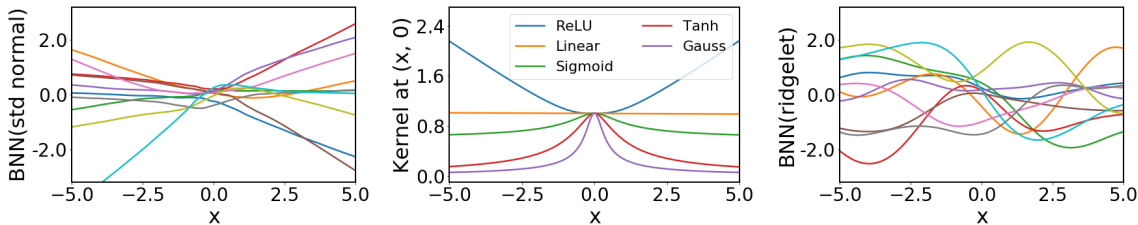


Figure 1: *Lifting the parameter prior distribution of a Bayesian neural network (BNN) to the output space of the network. Left:* Realisations from a BNN, with the ReLU activation function and independent standard Gaussian distributions on the weights and bias parameters. *Middle:* The covariance function of a BNN, as the activation function is varied over ReLU, linear, sigmoid, hyperbolic tangent and Gaussian. *Right:* Realisations from a BNN endowed with a ridgelet prior, which is constructed to approximate a GP with covariance  $k(\mathbf{x}, \mathbf{y}) = \sigma^2 \exp(-\frac{1}{2l^2} \|\mathbf{x} - \mathbf{y}\|_2^2)$  with  $\sigma = 1.0$ ,  $l = 1.75$ . [In all cases one hidden layer was used.]

Several deep connections between BNNs and GPs have been exposed (Rasmussen and Williams, 2006; Stein, 1999; Kristiadi et al., 2020; Khan et al., 2019; Adlam et al., 2021). The first detailed connection between BNNs and GPs was made by Neal (1995). Let  $[\mathbf{w}_i^l]_j := w_{i,j}^l$ . In the case of a shallow BNN with  $L = 1$ , assume that each of the weights  $\mathbf{w}_1^0$ ,  $\mathbf{w}_i^0$  and biases  $b_i^0$  are *a priori* independent, each with mean 0 and with finite second moments  $\sigma_{\mathbf{w}_1^0}^2, \sigma_{\mathbf{w}_i^0}^2, \sigma_b^2$ , respectively, where  $\sigma_{\mathbf{w}_1^0}^2 = \sigma^2/N_1$  for some fixed  $\sigma > 0$ . To improve presentation, let  $f(\mathbf{x}) := f(\mathbf{x}, \theta)$ , so that  $\theta$  is implicit, and let  $\mathbb{E}$  denote expectation with respect to  $\theta \sim \mathbb{P}$ . A well-known result from Neal (1995) is that, according to the central limit theorem, the BNN converges asymptotically (as  $N_1 \rightarrow \infty$ ) to a zero-mean GP with covariance function

$$k^1(\mathbf{x}, \mathbf{x}') := \mathbb{E}[f(\mathbf{x})f(\mathbf{x}')] = \sigma^2 \mathbb{E} [\phi(\mathbf{w}_1^0 \cdot \mathbf{x} + b_1^0) \phi(\mathbf{w}_1^0 \cdot \mathbf{x}' + b_1^0)] + \sigma_b^2, \quad (2)$$

Analytical forms of the GP covariance were obtained for several activation functions  $\phi$ , such as the ReLU and Gaussian error functions in, Lee et al. (2018); Williams (1998); Yang and Salman (2020). Furthermore, similar results were obtained more recently for neural networks with multiple hidden layers in Lee et al. (2018); Matthews et al. (2018); Novak et al. (2019); Garriga-Alonso et al. (2019). Placing independent priors with the same second moments  $\sigma^2/N_{l-1}$  and  $\sigma_b^2$  on the weights and biases of the  $l^{\text{th}}$  layer, and taking  $N_1 \rightarrow \infty, N_2 \rightarrow \infty, \dots$  in succession, it can be shown that the  $l^{\text{th}}$  layer of this BNN convergences to a zero mean GP with covariance:

$$k^l(\mathbf{x}, \mathbf{x}') = \sigma^2 \mathbb{E}_{z_i^{l-1} \sim \mathcal{GP}(0, k^{l-1})} [\phi(z_i^{l-1}(\mathbf{x})) \phi(z_i^{l-1}(\mathbf{x}'))] + \sigma_b^2. \quad (3)$$

Of course, the discussion of this section is informal only and we refer the reader to the original references for full and precise detail.

The identification of limiting forms of covariance function allows us to investigate whether such priors are suitable for performing uncertainty quantification in real-world

tasks. Unfortunately, the answer is often “no”. One reason, which has been demonstrated empirically by multiple authors (Hafner et al., 2020; Yang et al., 2019; Yao et al., 2019), is that BNNs can have poor out-of-sample performance. Typically one would want a covariance function to have a notion of locality, so that  $k(\mathbf{x}, \mathbf{x}')$  decays sufficiently rapidly as  $\mathbf{x}$  and  $\mathbf{x}'$  become distant from each other. This ensures that when predictions are made for a location  $\mathbf{x}$  that is far from the training dataset, the predictive variance is appropriately increased. However, as exemplified in Figure 1, the covariance structure of a BNN need not be local. Even the use of a light-tailed Gaussian activation function still has a possibility of leading to a covariance model that is non-local. These existing studies (Hafner et al., 2020; Yang et al., 2019; Yao et al., 2019) illustrate the difficulties of inducing a meaningful prior on the output space of the neural network when operating at the level of the parameters  $\theta$  of the network.

### 3. Methods

In this section the ridgelet approach to prior specification is presented. The approach relies on the classical ridgelet transform, which is briefly introduced in Section 3.1. Then, in Section 3.2, we describe how a ridgelet prior is constructed.

#### 3.1 The Ridgelet Transform

The *ridgelet transform* (Candès, 1998; Murata, 1996; Sonoda and Murata, 2017) was developed in the context of harmonic analysis in the 1990s (Barron, 1993; Jones, 1992; Leshno et al., 1993; Murata, 1996; Kurková and Sanguinetti, 2001) and has received recent interest as a tool for the theoretical analysis of neural networks (Sonoda and Murata, 2017; Bach, 2018; Sonoda et al., 2018; Ongie et al., 2020). In this section we provide a brief and informal description of the ridgelet transform, deferring all mathematical details until Section 4. To this end, let  $\hat{f}$  denote the Fourier transform of a function  $f$ , and let  $\bar{z}$  denote the complex conjugate of  $z \in \mathbb{C}$ . Given an activation function  $\phi : \mathbb{R} \rightarrow \mathbb{R}$ , suppose we have a corresponding function  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  such that the relationship

$$(2\pi)^{\frac{d}{2}} \int_{\mathbb{R}} |\xi|^{-d} \overline{\hat{\psi}(\xi)} \hat{\phi}(\xi) d\xi = 1$$

is satisfied. Such a function  $\psi$  is available in closed form for many of the activation functions  $\phi$  that are commonly used in neural networks; examples can be found in Table 1 in Section 4.2. Then, under regularity conditions detailed in Section 4, the ridgelet transform of a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is defined as

$$R[f](\mathbf{w}, b) := \int_{\mathbb{R}^d} \psi(\mathbf{w} \cdot \mathbf{x} + b) f(\mathbf{x}) d\mathbf{x} \tag{4}$$

for  $\mathbf{w} \in \mathbb{R}^d$ , and  $b \in \mathbb{R}$ , and the dual ridgelet transform of a function  $\tau : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}$  is defined as

$$R^*[\tau](\mathbf{x}) := \int_{\mathbb{R}^{d+1}} \phi(\mathbf{w} \cdot \mathbf{x} + b) \tau(\mathbf{w}, b) d\mathbf{w} db \tag{5}$$

for  $\mathbf{x} \in \mathbb{R}^d$ . There are two main properties of the ridgelet transform that we exploit in this work. First, a discretisation of (5) using a cubature method gives rise to an expression

closely related to one layer of a neural network; c.f. (1). Second, under regularity conditions, the dual ridgelet transform works as the pseudo-inverse of the ridgelet transform, meaning that  $(R^*R)[f] = f$  whenever the left hand side is defined. Next we explain how these two properties of the ridgelet transform will be used.

### 3.2 The Ridgelet Prior

In this section our proposed ridgelet prior is presented. Our starting point is a Gaussian stochastic process  $\mathcal{GP}(m, k)$  and we aim to construct a probability distribution for the parameters  $\theta$  of a neural network  $f(\mathbf{x}, \theta)$  in (1) such that the stochastic process  $\theta \mapsto f(\cdot, \theta)$  closely approximates the GP, in a sense yet to be defined.

The construction proceeds in three elementary steps. The first step makes use of the property that discretisation of  $R^*$  in (5) using a cubature method (i.e. a linear combination of function values) gives rise to a neural network. To see this, let us abstractly denote by  $\tilde{R}$  and  $\tilde{R}^*$  approximations of  $R$  and  $R^*$  obtained using a cubature method with  $D$  nodes:

$$R[f](\mathbf{w}, b) \approx \tilde{R}[f](\mathbf{w}, b) := \sum_{j=1}^D u_j \psi(\mathbf{w} \cdot \mathbf{x}_j + b) f(\mathbf{x}_j) \quad (6)$$

$$R^*[\tau](\mathbf{x}) \approx \tilde{R}^*[\tau](\mathbf{x}) := \sum_{i=1}^{N_1} v_i \phi(\mathbf{w}_i^0 \cdot \mathbf{x} + b_i^0) \tau(\mathbf{w}_i, b_i) \quad (7)$$

where  $(\mathbf{x}_j, u_j)_{j=1}^D \subset \mathbb{R}^d \times \mathbb{R}$  and  $((\mathbf{w}_i^0, b_i^0), v_i)_{i=1}^{N_1} \subset \mathbb{R}^{d+1} \times \mathbb{R}$  are the cubature nodes and weights employed respectively in (6) and (7). The specification of suitable cubature nodes and weights will be addressed in Section 4, but for now we assume that they have been specified. It is clear that (7) closely resembles one layer of a neural network; c.f. (1).

The second step makes use of the fact that  $(R^*R)[f] = f$ , which suggests that we may approximate a function  $f$  using the discretised ridgelet transform and its dual

$$\begin{aligned} (\tilde{R}^* \tilde{R})[f](\mathbf{x}) &= \sum_{i=1}^{N_1} v_i \phi(\mathbf{w}_i^0 \cdot \mathbf{x} + b_i^0) \left[ \sum_{j=1}^D u_j \psi(\mathbf{w}_i^0 \cdot \mathbf{x}_j + b_i^0) f(\mathbf{x}_j) \right] \\ &= \sum_{i=1}^{N_1} \underbrace{\sum_{j=1}^D [v_i u_j \psi(\mathbf{w}_i^0 \cdot \mathbf{x}_j + b_i^0) f(\mathbf{x}_j)]}_{=: w_{1,i}^1} \phi(\mathbf{w}_i^0 \cdot \mathbf{x} + b_i^0) \end{aligned} \quad (8)$$

where the coefficients  $\mathbf{w}_1^1 = (w_{1,1}^1, \dots, w_{1,N_1}^1)^\top$  depend explicitly on the function  $f$  being approximated. Thus  $\tilde{R}^* \tilde{R}$  is a linear operator that returns a neural network approximation to each function  $f$  provided as input.

The third and final step is to compute the pushforward of  $\mathcal{GP}(m, k)$  through the linear operator  $\tilde{R}^* \tilde{R}$ , in order to obtain a probability distribution over the coefficients  $\mathbf{w}_1^1$  of the neural network in (8). Let  $[\mathbf{f}]_i := f(\mathbf{x}_i)$ ,  $[\mathbf{m}]_i := m(\mathbf{x}_i)$ ,  $[\mathbf{K}]_{i,j} := k(\mathbf{x}_i, \mathbf{x}_j)$  and  $[\Psi^0]_{i,j} := v_i u_j \psi(\mathbf{w}_i^0 \cdot \mathbf{x}_j + b_i^0)$  so that  $\mathbf{f}, \mathbf{m} \in \mathbb{R}^D$ ,  $\mathbf{K} \in \mathbb{R}^{D \times D}$  and  $\Psi^0 \in \mathbb{R}^{N_1 \times D}$ . If



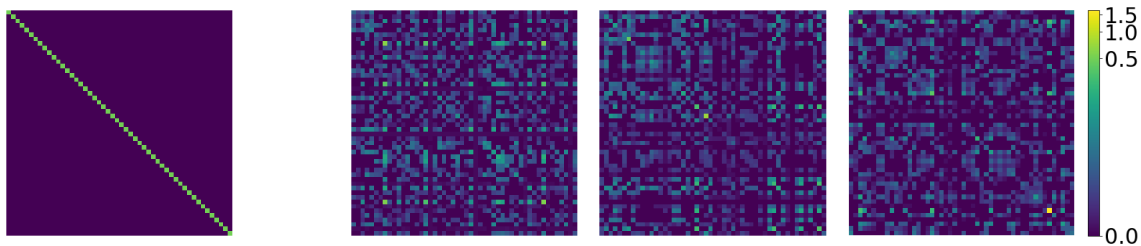


Figure 2: *Example of the covariance matrix  $\Psi^{l-1}\mathbf{K}(\Psi^{l-1})^\top$  of the ridgelet prior for  $l = 1$  and  $N_1 = 50$ . Left: The covariance matrix of independent standard Gaussian distribution prior on  $\mathbb{R}^{N_1}$  for comparison. Right: The covariance matrix  $\Psi^0\mathbf{K}(\Psi^0)^\top$  computed from 3 independent realisations of  $\{\mathbf{w}_i^0, b_i^0\}_{i=1}^{50}$  and from  $\mathbf{K}$  of a Gaussian covariance function.*

$f \sim \mathcal{GP}(m, k)$  then it follows immediately that  $\mathbf{w}_1^1 = \Psi^0 \mathbf{f}$  is a Gaussian random vector with

$$\begin{aligned} \mathbb{E}[\mathbf{w}_1^1] &= \Psi^0 \mathbf{m} \\ \mathbb{E}[(\mathbf{w}_1^1 - \mathbb{E}[\mathbf{w}_1^1])(\mathbf{w}_1^1 - \mathbb{E}[\mathbf{w}_1^1])^\top] &= \Psi^0 \mathbf{K}(\Psi^0)^\top. \end{aligned}$$

To arrive at a prior for a general neural network of the form (1) we apply this construction recursively, starting from the input layer and working up towards the output layer. The dimension of the cubature rule  $((\mathbf{w}_i^{l-1}, b_i^{l-1}), v_i^{l-1})_{i=1}^{N_l}$  used at level  $l - 1$  is required to equal  $N_l$  so that our discretised ridgelet transform inherits the same network architecture as in (1). Our notation is generalised to  $[\Psi^{l-1}]_{i,j} := v_i u_j \psi(\mathbf{w}_i^{l-1} \cdot \phi^{l-1}(\mathbf{x}_j) + b_i^{l-1})$  in order to indicate that this cubature rule with  $N_l$  elements was used to construct the matrix  $\Psi^{l-1}$ , where  $[\phi^{l-1}(\mathbf{x}_j)]_i := \phi(z_i^{l-1}(\mathbf{x}_j))$  and  $\phi^0(\mathbf{x}_j) = \mathbf{x}_j$ . Let  $N_{L+1}$  be the output dimension of the neural network in (1), that is  $N_{L+1} = 1$ . Our ridgelet prior can now be formally defined:

**Definition 1** (Ridgelet Prior). *Consider the neural network in (1). Given a mean function  $m : \mathbb{R}^d \rightarrow \mathbb{R}$  and a covariance function  $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ , a prior distribution is called a ridgelet prior if the weights  $\mathbf{w}_i^l$  at level  $l$  depend on the weights and biases at level  $l - 1$  according to*

$$\mathbf{w}_i^l | \{(\mathbf{w}_r^{l-1}, b_r^{l-1}) : r = 1, \dots, N_l\} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\Psi^{l-1} \mathbf{m}, \Psi^{l-1} \mathbf{K}(\Psi^{l-1})^\top)$$

where  $i = 1, \dots, N_{l+1}$  and  $l = 1, \dots, L$ . To complete the prior specification, the bias parameters  $b_r^{l-1}$  at all layers and the weights  $\mathbf{w}_i^0$  at the input layer are required to be independent and distributed as  $b_i^{l-1} \sim \mathcal{N}(0, \sigma_b^2)$  and  $\mathbf{w}_i^0 \sim \mathcal{N}(0, \sigma_w^2 \mathbf{I}_{d \times d})$ .

Several initial remarks are in order:

**Remark 1.** *The dependence of the distribution for the weights  $\mathbf{w}_i^l$  on the previous layer's weights  $\mathbf{w}_r^{l-1}$  and the biases  $b_r^{l-1}$ ,  $r = 1, \dots, N_l$ , is an important feature of our construction and seems essential if we aim to approximate a user-specified covariance model. This dependence is illustrated in Figure 2.*

**Remark 2.** Let  $[\phi^0(\mathbf{x})]_i := \phi(\mathbf{w}_i^0 \cdot \mathbf{x} + b_i^0)$ . In the comparison with (2), the covariance of a BNN with one hidden layer and endowed with our ridgelet prior takes the form

$$\mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] = \mathbb{E}[\phi^0(\mathbf{x})^\top \Psi^0 \mathbf{K}(\Psi^0)^\top \phi^0(\mathbf{x}')] \quad (9)$$

where the expectation on the right hand side is with respect to the random weights  $\mathbf{w}_i^0$  and biases  $b_i^0$ . In Section 4 it is shown that, in an appropriate limit, the expression in (9) converges to  $k(\mathbf{x}, \mathbf{x}')$ , the covariance model that we aimed to approximate at the outset.

**Remark 3.** To limit scope we consider a standard feed-forward network architecture, but it is possible in principle to extend the ridgelet prior to other types of network. For convolutional neural networks (CNN), we note that convolutional layers can be seen as feed-forward layers with a sparse weight matrix. In this case one could consider taking the ridgelet prior as a starting point and then condition on specific weights being exactly zero, in order to produce identical sparsity structure to a CNN (see for example Section 2.1 of Garriga-Alonso et al., 2019). For residual neural networks (ResNets), the predictor vector  $\phi^l$  can be augmented to include any skip connections that may be present. However, our analysis in Section 4 focuses on the simple case of a single hidden layer, for which both CNNs and ResNets reduce to a standard feed-forward neural network.

This completes our definition of the ridgelet prior. An illustration is provided in Figure 3, the full details of which are reserved for Section 6.1. It can be seen that as  $N$ , the number of hidden units in the BNN, is increased the samples from the BNN begin to resemble, in a statistical sense, samples from the target GP. In the case of multiple hidden layers, a larger number  $N$  of hidden units appear to be required to achieve a similar degree of approximation to the GP. Next we present our theoretical analysis, which considers only the case of one hidden layer, and is the principal contribution of the paper.

## 4. Theoretical Assessment

This section presents our theoretical analysis of the ridgelet prior in the setting of a single hidden layer ( $L = 1$ ). The main result is Theorem 1, which establishes an explicit error bound for approximation of a GP prior using the ridgelet prior in a BNN. The bound is non-asymptotic and valid for networks with finite numbers of units, which is an important point of distinction from earlier work (Neal, 1995; Williams, 1998). First, in Section 4.1, we introduce definitions and notation required for the ridgelet transform to be rigorously studied. In Section 4.2 we introduce a finite-bandwidth ridgelet transform, with the error between the finite-bandwidth and original ridgelet transforms being analysed in Appendix A.1. Section 4.3 applies these finite-bandwidth results to perform a theoretical analysis of the ridgelet prior for a BNN.

**Notation:** The following notation will be used. For an open set  $\mathcal{X} \subseteq \mathbb{R}^d$  and a Borel measure  $\mu$  on  $\mathcal{X}$ , let  $L^p(\mathcal{X}, \mu)$  denote the set of functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  such that  $\|f\|_{L^p(\mathcal{X}, \mu)} < \infty$  where

$$\|f\|_{L^p(\mathcal{X}, \mu)} := \begin{cases} (\int_{\mathcal{X}} |f(x)|^p d\mu(x))^{1/p} & 1 \leq p < \infty \\ \text{ess sup}_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x})| & p = \infty. \end{cases}$$

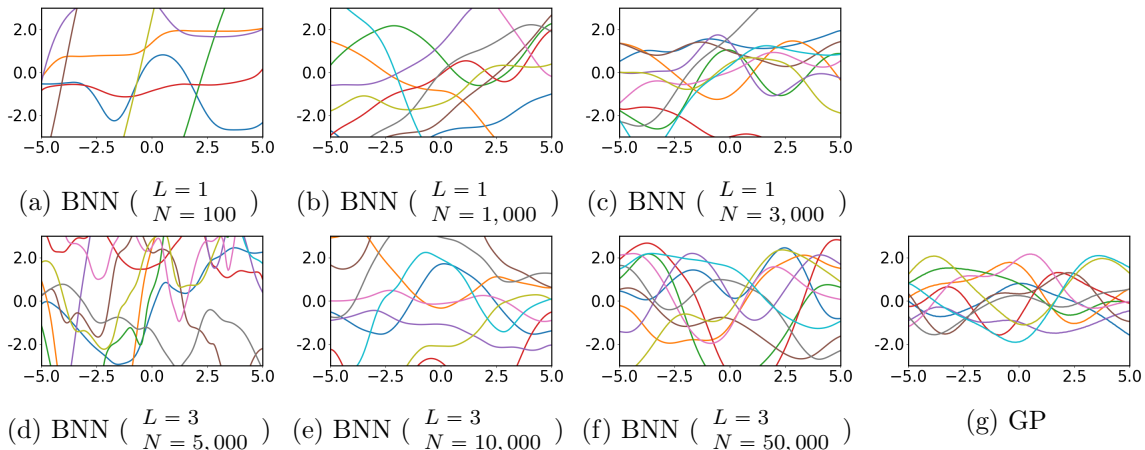


Figure 3: Sample paths from a Bayesian neural network (BNN) equipped with a ridgelet prior. Here we examine the effect of increasing the number  $N$  of hidden units in a BNN with  $L = 1$  hidden layer from (a) 100, to (b) 1,000 and (c) 3,000, and in a BNN with  $L = 3$  hidden layers from (d) 5,000, to (e) 10,000 and (f) 50,000. In (g) sample paths are shown from the target Gaussian process (GP) prior. Full details of this example are reserved for Section 6.1.

If  $\mu$  is the Lebesgue measure on  $\mathcal{X}$ , we use the shorthand  $L^p(\mathcal{X})$  for  $L^p(\mathcal{X}, \mu)$  and furthermore we let  $L^p_{\text{loc}}(\mathcal{X})$  denote the set of function  $f : \mathcal{X} \rightarrow \mathbb{R}$  such that  $\|f\|_{L^p(K)}$  exists and is finite on all compact  $K \subseteq \mathcal{X}$ . Let  $C(\mathcal{X})$  denote the set of all continuous functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  and let  $C^1(\mathcal{X})$  denote the set of all functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  whose first order partial derivatives  $\partial_{x_i} f(\mathbf{x})$  exist and are continuous on  $\mathcal{X}$ . Denote by  $C^{1 \times 1}(\mathcal{X} \times \mathcal{X})$  the set of all functions  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  whose mixed first order derivatives  $\partial_{x_i} \partial_{y_j} h(\mathbf{x}, \mathbf{y})$  exist and are continuous on  $\mathcal{X} \times \mathcal{X}$ . For multinomials and higher-order derivatives, we employ multi-index notation  $\mathbf{x}^\alpha := x_1^{\alpha_1} \dots x_d^{\alpha_d}$ ,  $\partial^\alpha h(\mathbf{x}) := \partial_{x_1}^{\alpha_1} \dots \partial_{x_d}^{\alpha_d} h(\mathbf{x})$  and  $\partial^{\alpha, \beta} h(\mathbf{x}, \mathbf{y}) := \partial_{x_1}^{\alpha_1} \dots \partial_{x_d}^{\alpha_d} \partial_{y_1}^{\beta_1} \dots \partial_{y_d}^{\beta_d} h(\mathbf{x}, \mathbf{y})$ . A bivariate function  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is called *positive definite* if  $\sum_{i,j} a_i a_j k(\mathbf{x}_i, \mathbf{x}_j) > 0$  for all  $\mathbf{0} \neq (a_1, \dots, a_n) \in \mathbb{R}^n$  and all distinct  $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$ .

#### 4.1 Regularity of the Activation Function

In this section we outline our regularity assumptions on the activation function  $\phi$ . To do this we first recall the classical Fourier transform and its generalisations, all for real-valued functions on  $\mathbb{R}^d$ .

**Fourier transform on  $L^1(\mathbb{R}^d)$  and  $L^2(\mathbb{R}^d)$ :** The Fourier transform of  $f \in L^1(\mathbb{R}^d)$  is defined by  $\hat{f}(\boldsymbol{\xi}) := (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} f(\mathbf{x}) \exp(-i\boldsymbol{\xi} \cdot \mathbf{x}) d\mathbf{x}$  for each  $\boldsymbol{\xi} \in \mathbb{R}^d$ . The Fourier transform of  $f \in L^2(\mathbb{R}^d)$  is formally defined as a limit of a sequence  $(\hat{f}_n)_{n \in \mathbb{N}}$  where the  $f_n \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$  and  $f_n \rightarrow f$  in  $L^2(\mathbb{R}^d)$  (see e.g. Grafakos, 2000, p.113-114).

The image of the Fourier transform on  $L^1(\mathbb{R}^d)$  is not contained in  $L^1(\mathbb{R}^d)$ . However, there exists a subset of  $L^1(\mathbb{R}^d)$  on which the Fourier transform defines an automorphism. It

is convenient to work on this subset, which consists of so-called *Schwartz functions*, defined next.

**Schwartz functions:** An infinitely differentiable function  $f$  is called a *Schwartz function* if for any pair of multi-indices  $\alpha, \beta \in \mathbb{N}_0^d$  we have that  $C_{\alpha, \beta} := \sup_{\mathbf{x} \in \mathbb{R}^d} |\mathbf{x}^\alpha \partial^\beta f(\mathbf{x})| < \infty$ . The set of Schwartz functions is denoted by  $\mathcal{S}(\mathbb{R}^d)$  (Grafakos, 2000, p.105). Note that the Fourier transform on  $\mathcal{S}(\mathbb{R}^d)$  is well-defined as  $\mathcal{S}(\mathbb{R}^d) \subset L^1(\mathbb{R}^d)$ . Moreover, the Fourier transform is a homeomorphism from  $\mathcal{S}(\mathbb{R}^d)$  onto itself (Grafakos, 2000, p.113).

The most commonly encountered activation functions  $\phi$  are not elements of  $L^1(\mathbb{R})$  and we therefore also require the notion of a generalised Fourier transform:

**Generalised Fourier transform:** Let  $\mathcal{S}_m(\mathbb{R}^d)$  be the vector space of functions  $f \in \mathcal{S}(\mathbb{R}^d)$  that satisfy  $f(\mathbf{x}) = \mathcal{O}(\|\mathbf{x}\|^m)$  for  $\mathbf{x} \rightarrow 0$ . For  $k \in \mathbb{N}_0$ , let  $t : \mathbb{R}^d \rightarrow \mathbb{R}$  be any continuous function of at most polynomial growth of order  $k$ , meaning that  $\sup_{\mathbf{x} \in \mathbb{R}^d} |t(\mathbf{x})|/(1 + \|\mathbf{x}\|^k) < \infty$ . A measurable function  $\hat{t} \in L^2_{\text{loc}}(\mathbb{R}^d \setminus \{\mathbf{0}\})$  is called a *generalised Fourier transform* of  $t$  if there exists an integer  $m \in \mathbb{N}_0$  such that  $\int_{\mathbb{R}^d} \hat{t}(\mathbf{w})f(\mathbf{w})d\mathbf{w} = \int_{\mathbb{R}^d} t(\mathbf{x})\hat{f}(\mathbf{x})d\mathbf{x}$  for all  $f \in \mathcal{S}_{2m}(\mathbb{R}^d)$  (Wendland, 2005, p.103). The set of all continuous functions of at most polynomial growth of order  $k$  that admit a generalized Fourier transform will be denoted  $C_k^*(\mathbb{R}^d)$ .

The generalized Fourier transform can be computed for activation functions  $\phi$  that are typically used in a neural network, for which classical Fourier transforms are not well-defined:

**Example 1** (Generalised Fourier transform of the ReLU function). *Let  $\phi(x) := \max(0, x)$ , then  $\phi \in C_1^*(\mathbb{R})$ . Although  $\phi$  is not an element of  $L^1(\mathbb{R}^d)$ , it has polynomial growth and admits a generalised Fourier transform  $\hat{\phi}(w) = -(\sqrt{2\pi}w^2)^{-1}$ .*

**Example 2** (Generalised Fourier transform of the tanh function). *Let  $\phi(x) := \tanh(x)$ , then  $\phi \in C_0^*(\mathbb{R})$ . Likewise,  $\phi$  admits a generalised Fourier transform  $\hat{\phi}(w) = -i\sqrt{\frac{\pi}{2}}\text{csch}(\frac{\pi}{2}w)$  where  $\text{csch}$  is the hyperbolic cosecant.*

In order to present our theoretical results, we will make the following assumptions on the activation function  $\phi$  that defines the ridgelet transform:

**Assumption 1** (Activation function). *The activation function  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  satisfies*

1.  $\phi \in C_0^*(\mathbb{R})$ ,

and there exists a function  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  such that

2.  $\psi \in \mathcal{S}(\mathbb{R})$ ,
3.  $(2\pi)^{\frac{d}{2}} \int_{\mathbb{R}} |\xi|^{-d} \overline{\hat{\psi}(\xi)} \hat{\phi}(\xi) d\xi = 1$ ,
4.  $\int_{\mathbb{R}} |\xi|^{-d-2} |\overline{\hat{\psi}(\xi)} \hat{\phi}(\xi)| d\xi < \infty$ .

The boundedness assumption  $\phi \in C_0^*(\mathbb{R})$  rules out some commonly used activation functions, such as ReLU, but enables stronger convergence results to be obtained. However, our analysis is also able to handle  $\phi \in C_1^*(\mathbb{R})$ . For presentational purposes we present results under the assumption  $\phi \in C_0^*(\mathbb{R})$  in the main text and, in Appendix A.3 we present theoretical results for the unbounded setting  $\phi \in C_1^*(\mathbb{R})$ . Parts 2 and 3 of Assumption 1 are the standard assumptions for the ridgelet transform (Candès, 1998; Murata, 1996); several examples of  $(\phi, \psi)$  pairs are shown in Table 1.

Next we turn our attention to proving novel and general results about the ridgelet transform, under Assumption 1.

Activation $\phi$	Associated Function $\psi$
$\tanh(z)$	$\frac{d^{d-r+2}}{dz^{d-r+2}} \left[ \exp\left(-\frac{z^2}{2}\right) \sin\left(\frac{\pi z}{2}\right) \right] \times 2^{-\frac{d+r}{2}} \pi^{-\frac{d-r+2}{2}} \exp\left(\frac{\pi^2}{8}\right)$
$(1 + \exp(-z))^{-1}$	$\frac{d^{d-r+2}}{dz^{d-r+2}} \left[ \exp\left(-\frac{z^2}{2}\right) \sin(\pi z) \right] \times 2^{-\frac{d+r}{2}} \pi^{-\frac{d-r+2}{2}} \exp\left(\frac{\pi^2}{2}\right)$
$\exp\left(-\frac{z^2}{2}\right)$	$\frac{d^{d+r}}{dz^{d+r}} \exp\left(-\frac{z^2}{2}\right) \times 2^{-\frac{d-2r}{2}} \pi^{-\frac{d+r}{2}}$
$\max(0, z)$	$\frac{d^{d+r+2}}{dz^{d+r+2}} \exp\left(-\frac{z^2}{2}\right) \times \left(-2^{-\frac{d}{2}} \pi^{-\frac{d-2r+1}{2}}\right)$

Table 1: Examples of functions  $\phi, \psi$  that satisfy the regularity assumptions used in this work where  $r = (d \bmod 2)$ .

## 4.2 A Finite-Bandwidth Ridgelet Transform

Our aim in this section is to approximate the dual ridgelet transform in (5) with a *finite-bandwidth* transform, meaning that the Lebesgue measure  $\lambda(\mathbf{w}, b)$  is approximated using a finite measure  $\lambda_\sigma(\mathbf{w}, b)$ . This will in turn enable cubature methods to be applied to discretise (5). In anticipation of our analysis of the ridgelet prior in Definition 1, the finite measure we consider is

$$\lambda_\sigma(\mathbf{w}, b) := Z \times \frac{1}{(2\pi\sigma_{\mathbf{w}}^2)^{\frac{d}{2}}} \exp\left(-\frac{\|\mathbf{w}\|^2}{2\sigma_{\mathbf{w}}^2}\right) \times \frac{1}{(2\pi\sigma_b^2)^{\frac{1}{2}}} \exp\left(-\frac{b^2}{2\sigma_b^2}\right).$$

where it will be convenient to set  $Z := (2\pi)^{\frac{1}{2}} \sigma_{\mathbf{w}}^d \sigma_b$  indicating the total measure assigned by  $\lambda_\sigma$  to  $\mathbb{R}^{d+1}$ . It is possible to generalise our analysis beyond this Gaussian case, and other choices for  $\lambda_\sigma$  are analysed in Appendix A.1.

**Definition 2** (Finite-bandwidth ridgelet transform). *In the setting of Assumption 1, the ridgelet transform  $R$  and the finite-bandwidth approximation  $R_\sigma^*$  of its dual  $R^*$  are defined pointwise as*

$$R[f](\mathbf{w}, b) := \int_{\mathbb{R}^d} f(\mathbf{x}) \psi(\mathbf{w} \cdot \mathbf{x} + b) d\mathbf{x}, \quad (10)$$

$$R_\sigma^*[\tau](\mathbf{x}) := \int_{\mathbb{R}^{d+1}} \tau(\mathbf{w}, b) \phi(\mathbf{w} \cdot \mathbf{x} + b) d\lambda_\sigma(\mathbf{w}, b) \quad (11)$$

for all  $f \in L^1(\mathbb{R}^d)$ ,  $\tau \in L^1(\mathbb{R}^{d+1}, \lambda_\sigma)$ ,  $\mathbf{x} \in \mathbb{R}^d$ ,  $\mathbf{w} \in \mathbb{R}^d$ , and  $b \in \mathbb{R}$ .

The integral transforms in (10) and (11) are indeed well-defined: For any  $f \in L^1(\mathbb{R}^d)$ , the boundedness of  $\psi \in \mathcal{S}(\mathbb{R})$  guarantees  $R[f] \in L^\infty(\mathbb{R}^{d+1})$ . Similarly for any  $\tau \in L^1(\mathbb{R}^{d+1}, \lambda_\sigma)$ , the boundedness of  $\phi \in C_0^*(\mathbb{R})$  guarantees  $R_\sigma^*[\tau] \in L^\infty(\mathbb{R}^d)$ . Recall that a discussion on relaxation of the boundedness assumption  $\phi \in C_0^*(\mathbb{R})$  is reserved for Appendix A.3.

The classical ridgelet transform in the sense of Murata (1996) corresponds to the limit  $\sigma_{\mathbf{w}}, \sigma_b \rightarrow \infty$ , where the measure  $\lambda_\sigma$  becomes flat and  $R_\sigma^*$  coincides with  $R^*$ . An original contribution of our work, which may be of more general interest, is to study approximation properties of the reconstruction operator  $R_\sigma^* R$  when a finite measure  $\lambda_\sigma$  is used. Intuitively,  $(R_\sigma^* R)$  ought to converge to an identity operator in the limit  $\sigma_{\mathbf{w}}, \sigma_b \rightarrow \infty$ ; an important contribution of this work is to provide an explicit, non-asymptotic approximation error bound in Appendix A.1. This analysis, which connects existing theory of the ridgelet transform

to practical applications, may therefore be useful in the aforementioned applications of harmonic analysis, as well as in related areas, such as image analysis/denoising, where the ridgelet transform is sometimes employed as an alternative to the wavelet transform (Do and Vetterli, 2003; AlZu’bi et al., 2011).

### 4.3 Convergence of the Finite-Bandwidth Ridgelet Prior

In this section, we build on Theorem 2 to obtain theoretical guarantees for our BNN prior. We present Theorem 1, which establishes an explicit error bound for approximation of a GP prior using the ridgelet prior in a BNN.

Recall from Section 3 that the ridgelet prior was derived from discretisation of the classical ridgelet transform in (4) and (5). The starting point of our analysis is to consider how the finite-bandwidth ridgelet transform in (10) and (11) can be discretised. To this end, we view the discretisation of  $R_\sigma^* R$  as an operator  $I_{\sigma,D,N} : C(\mathbb{R}^d) \rightarrow C(\mathbb{R}^d)$  by

$$I_{\sigma,D,N}[f](\mathbf{x}) := \sum_{i=1}^N v_i \left( \sum_{j=1}^D u_j f(\mathbf{x}_j) \psi(\mathbf{w}_i \cdot \mathbf{x}_j + b_i) \right) \phi(\mathbf{w}_i \cdot \mathbf{x} + b_i) \quad (12)$$

where  $\{\mathbf{x}_j, u_j\}_{j=1}^D$  and  $\{(\mathbf{w}_i, b_i), v_i\}_{i=1}^N$  are the cubature nodes and weights used, respectively, to discretise the integrals (10) and (11). Informally,  $I_{\sigma,D,N}$  ought to converge in some sense to the identity operator in an appropriate limit involving  $\sigma, D, N \rightarrow \infty$ ; this will be made precise in the sequel. The weights  $\mathbf{w}_i$  and biases  $b_i$  can be identified with  $\mathbf{w}_i^0$  and  $b_i^0$  in the ridgelet prior, as can  $N_1$  be identified with  $N$ . Conditional on  $\{\mathbf{w}_i, b_i\}_{i=1}^N$ , a draw from the BNN  $f(\mathbf{x}, \theta) = \sum_{i=1}^{N_1} w_{1,i}^1 \phi(\mathbf{w}_i^0 \cdot \mathbf{x} + b_i^0)$  equipped with the ridgelet prior is equal in distribution to  $I_{\sigma,D,N}[f]$  when  $f$  is drawn from  $\mathcal{GP}(m, k)$ .

Our analysis exploits properties of the cubature methods defined in Assumption 2 and Assumption 3 below. To rigorously introduce these cubature methods, let  $\mathcal{X}$  be a bounded subset of  $\mathbb{R}^d$ ; without loss of generality we assume  $\mathcal{X}$  is contained in the interior of the hypercube  $[-S, S]^d$  for some constant  $S > 0$ . It follows that there exists an infinitely differentiable function (a ‘‘mollifier’’)  $\mathbb{1}(\cdot)$  with the properties that  $\mathbb{1}(\mathbf{x}) \in [0, 1]$ ,  $\mathbb{1}(\mathbf{x}) = 1$  if  $\mathbf{x} \in \mathcal{X}$  and  $\mathbb{1}(\mathbf{x}) = 0$  if  $\mathbf{x} \notin [-S, S]^d$ . Indeed, letting  $a := 1 - \inf\{\|\mathbf{x} - \mathbf{y}\| : \mathbf{x} \in \mathcal{X}, \mathbf{y} \notin [-S, S]^d\}$  and

$$g(t) := \begin{cases} e^{-\frac{1}{t}} & \text{for } t > 0 \\ 0 & \text{for } t \leq 0 \end{cases} \quad \text{and} \quad h(t) := \frac{g(t)}{g(t) + g(1-t)}$$

then the function

$$\mathbb{1}(\mathbf{x}) := \prod_{i=1}^d \left[ 1 - h\left(\frac{x_i^2 - a^2}{1 - a^2}\right) \right] \quad (13)$$

satisfies the desired properties with  $\mathbb{1}(\cdot)$  (Tu, 2010, p.141-143). The function  $\mathbb{1}(\cdot)$  is used in our theoretical analysis to restrict, via multiplication, the support of function  $f$  from  $\mathbb{R}^d$  to  $[-S, S]^d$  without changing  $f$  on  $\mathcal{X}$  and without global smoothness on  $\mathbb{R}^d$  being lost; see Figure 4.

**Assumption 2** (Discretisation of  $R$ ). *The cubature nodes  $\{\mathbf{x}_j\}_{j=1}^D$  are a regular grid on  $[-S, S]^d$ , corresponding to a Cartesian product of left endpoint rules ((2.1.2) of Davis and Rabinowitz, 1984), and the cubature weights are  $u_j := (2S)^d \mathbb{1}(\mathbf{x}_j)/D$  for all  $j = 1, \dots, D$ .*

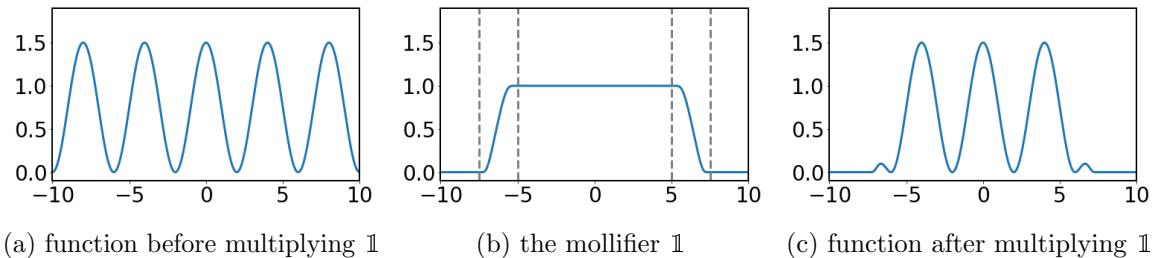


Figure 4: Illustrating the role of a mollifier: (a) A function  $f(\cdot)$  defined on  $\mathbb{R}$ , to be mollified. (b) Here  $\mathbb{1}$  is mollifier that is equal to 1 on  $\mathcal{X} = (-5, 5)$  and equal to 0 outside  $[-S, S]$  with  $S = 7.5$ . (c) The product  $f \cdot \mathbb{1}$  is equal to  $f$  on  $\mathcal{X}$  and has a compact support on  $[-S, S]$ .

The use of a Cartesian grid cubature method is not crucial to our analysis and other methods could be considered. However, we found that the relatively weak assumptions required to ensure convergence of a cubature method based on a Cartesian grid were convenient to check. One could consider adapting the cubature method to exploit additional smoothness that may be present in the integrand, but we did not pursue that in this work.

**Assumption 3** (Discretisation of  $R_\sigma^*$ ). *The cubature nodes  $\{\mathbf{w}_i, b_i\}_{i=1}^N$  are independently sampled from  $\mathcal{N}(0, \sigma_w^2 \mathbf{I}_{d \times d}) \times \mathcal{N}(0, \sigma_b^2)$  and  $v_i := Z/N$  for all  $i = 1, \dots, N$ .*

The use of a Monte Carlo cubature scheme in Assumption 3 ensures an exact identification between the ridgelet prior, where the  $(\mathbf{w}_i, b_i)$  are *a priori* independent, and a cubature method. From the approximation theoretic viewpoint,  $I_{\sigma, D, N}$  is now a random operator and we must account for this randomness in our analysis; high probability bounds are used to this effect in Theorem 1.

Now we present our main result, which is a finite-sample-size error bound for the ridgelet prior as an approximation of a GP. The approximation error that we study is defined over  $\mathcal{X} \subset [-S, S]^d$  and to this end we introduce the notation

$$M_1^*(m) := \max_{|\alpha| \leq 1} \sup_{\mathbf{x} \in [-S, S]^d} |\partial^\alpha m(\mathbf{x})|, \quad M_1^*(k) := \max_{|\alpha|, |\beta| \leq 1} \sup_{\mathbf{x}, \mathbf{y} \in [-S, S]^d} |\partial_{\mathbf{x}}^\alpha \partial_{\mathbf{y}}^\beta k(\mathbf{x}, \mathbf{y})|.$$

In what follows we introduce a random variable  $f$  to represent the GP, and we assume that  $f$  is independent of the random variables  $\{\mathbf{w}_i, b_i\}_{i=1}^N$ ; all random variables are understood to be defined on a common underlying probability space whose expectation operator is denoted  $\mathbb{E}$ .

**Theorem 1.** *Consider a stochastic process  $f \sim \mathcal{GP}(m, k)$  with mean function  $m \in C^1(\mathbb{R}^d)$  and symmetric positive definite covariance function  $k \in C^{1 \times 1}(\mathbb{R}^d \times \mathbb{R}^d)$ . Let Assumptions 1, 2 and 3 hold. Further, assume  $\phi$  is  $L_\phi$ -Lipschitz continuous. With probability at least  $1 - \delta$ ,*

$$\begin{aligned} & \sup_{\mathbf{x} \in \mathcal{X}} \sqrt{\mathbb{E} \left[ (f(\mathbf{x}) - I_{\sigma, D, N} f(\mathbf{x}))^2 \mid \{\mathbf{w}_i, b_i\}_{i=1}^N \right]} \\ & \leq C \left( M_1^*(m) + \sqrt{M_1^*(k)} \right) \left\{ \frac{1}{\sigma_w} + \frac{\sigma_w^d (\sigma_w + 1)}{\sigma_b} + \frac{\sigma_b \sigma_w^d (\sigma_w + 1)}{D^{\frac{1}{d}}} + \frac{\sigma_b \sigma_w^d (\sigma_w + \sqrt{\log \delta^{-1}})}{\sqrt{N}} \right\} \end{aligned}$$

where  $C$  is a constant independent of  $m, k, \sigma_w, \sigma_b, D, N$  and  $\delta$ .

**Proof** The proof is established in Appendix A.2. ■

**Remark 4.** *Theorem 1 is analogous to universal approximation theorems for (non-Bayesian) neural networks (Cybenko, 1989; Leshno et al., 1993), since it implies that shallow BNNs are capable of approximating essentially any GP whose mean and covariance functions are continuously differentiable.*

This result establishes that the root-mean-square error between the GP and its approximating BNN will vanish uniformly over the bounded set  $\mathcal{X}$  as the bandwidth parameters  $\sigma_w, \sigma_b$  and the sizes  $N, D$  of the cubature rules are increased at appropriate relative rates, which can be read off from the bound in Theorem 1. In particular, the bandwidths  $\sigma_w$  and  $\sigma_b$  should be increased in such a manner that  $\sigma_w^d(\sigma_w + 1)/\sigma_b \rightarrow 0$  and the numbers  $D$  and  $N$  of cubature nodes should be increased in such a manner that the final two terms in the bound of Theorem 1 asymptotically vanish. In such a limit, the ridgelet prior provides a consistent approximation of the GP prior over the bounded set  $\mathcal{X}$ . The result holds with high probability  $(1 - \delta)$  with respect to randomness in the sampling of the weights and biases  $\{\mathbf{w}_i, b_i\}_{i=1}^N$ . From this concentration inequality, and in an appropriate limit of  $\sigma_w, \sigma_b, N$  and  $D$ , the almost sure convergence of the root-mean-square error also follows (from Borel-Cantelli).

This finite-sample-size error bound can be contrasted with earlier work that provided only asymptotic convergence results (Neal, 1995; Lee et al., 2018; Matthews et al., 2018).

**Remark 5.** *The rates of convergence serve as a proof-of-concept and are the usual rates that one would expect in the absence of anisotropy assumptions, with the curse of dimension appearing in the term  $D^{-1/d}$ . In other words, we are gated by the rate at which we fill the volume  $[-S, S]^d$  with cubature nodes  $\{\mathbf{x}_j\}_{j=1}^D$ . However, it may be the case that real applications possess additional regularity, such as a low “effective dimension”, that was not part of the assumptions used to obtain our bound. We therefore defer further discussion of the approximation error to the empirical assessment in Section 6.*

**Remark 6.** *Related result can also be obtained when the boundedness assumption on the activation function is relaxed from  $\phi \in C_0^*(\mathbb{R})$  to  $\phi \in C_1^*(\mathbb{R})$ . Details are reserved for Appendix A.3.*

This completes our analysis of the ridgelet prior for BNNs and our attention turns, in the next section, to the empirical performance of the ridgelet prior.

## 5. Implementation and Computation

In this section, practical aspects of the ridgelet prior are discussed. Since our main contributions are theoretical, this discussion is limited to a factual description of the complexity of the ridgelet prior (Section 5.1), a description of how computation was performed for the experiments reported in Section 6 (Section 5.2), and a brief discussion of the potential for further computational improvement (Section 5.3).



### 5.1 Complexity of the Ridgelet Prior

To evaluate the ridgelet prior density, consider the  $l$ th layer and its associated Gaussian factor  $\mathcal{N}(0, \boldsymbol{\Sigma}^l)$ . To evaluate the prior density three computations are required: matrix multiplication to compute  $\boldsymbol{\Sigma}^l = \boldsymbol{\Psi}^{l-1} \mathbf{K} (\boldsymbol{\Psi}^{l-1})^\top$ , inversion of  $\boldsymbol{\Sigma}^l$  and computation of the determinant of  $\boldsymbol{\Sigma}^l$ . The total computational complexity is therefore  $O(N^l D^2 + (N^l)^2 D + (N^l)^3)$ , which is cubic in  $N^l$ . However, to sample from  $\mathcal{N}(0, \boldsymbol{\Sigma}^l)$  the computation complexity can be reduced when  $D \ll N^l$ , since we can obtain a matrix square root of the form  $(\boldsymbol{\Sigma}^l)^{1/2} := \boldsymbol{\Psi}^{l-1} \mathbf{L}$  via a Cholesky decomposition  $\mathbf{K} = \mathbf{L} \mathbf{L}^\top$ . The complexity is then  $O(D^3 + N^l D)$ , which is linear in  $N^l$ .

### 5.2 Posterior Computation

For the subsequent experiments reported in Section 6.2, the posterior distribution over the parameters of the BNN is analytically intractable and must therefore be approximated. In this paper, we exploited conditional linearity to obtain accurate approximations to the posterior using a Monte Carlo method, which will now be described. All our experiments concern regression models of the form

$$\mathbf{y}^{(i)} = f(\mathbf{x}^{(i)}, \boldsymbol{\theta}) + \epsilon^{(i)}, \quad f(\mathbf{x}, \boldsymbol{\theta}) = \sum_{i=1}^N w_i^1 \phi(\mathbf{w}_i^0 \cdot \mathbf{x} + b_i^0), \quad \epsilon^{(i)} \sim \mathcal{N}(0, \sigma_\epsilon^2). \quad (14)$$

Letting  $\boldsymbol{\theta}^0 := ((w_1^0, b_1^0), \dots, (w_N^0, b_N^0))$  and  $\mathbf{w}^1 := (w_1^1, \dots, w_N^1)$ , the ridgelet prior takes the form

$$\mathbf{w}_i^0 \sim \mathcal{N}(0, \sigma_w^2 I_{d \times d}), \quad b_i^0 \sim \mathcal{N}(0, \sigma_b^2), \quad \mathbf{w}^1 | \boldsymbol{\theta}^0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where  $\boldsymbol{\Sigma} := \boldsymbol{\Psi}_0 \mathbf{K} \boldsymbol{\Psi}_0^\top$ . To simplify this discussion, we assume the mean function  $m$  of the target GP is  $m(\mathbf{x}) = 0$ ; if not then one can replace the responses  $y^{(i)}$  with  $y^{(i)} - m(\mathbf{x}^{(i)})$  in the sequel. Let  $[\mathbf{y}]_i = y^{(i)}$  and  $[\boldsymbol{\Phi}_0]_{i,j} := \phi(\mathbf{w}_j^0 \cdot \mathbf{x}^{(i)} + b_j^0)$ . Our main observation is that, conditional on  $\boldsymbol{\theta}^0$ , the model (14) is a linear regression whose coefficients  $\mathbf{w}^1$  can be analytically marginalised to obtain a Gaussian marginal (log-)likelihood

$$\log p(\mathbf{y} | \boldsymbol{\theta}^0) = C - \frac{1}{2} \log \det \boldsymbol{\Sigma}_* - \frac{1}{2} \mathbf{y}^\top \boldsymbol{\Sigma}_*^{-1} \mathbf{y}$$

where  $\boldsymbol{\Sigma}_* := \boldsymbol{\Phi}_0 \boldsymbol{\Sigma} \boldsymbol{\Phi}_0^\top + \sigma_\epsilon^2 I$  and  $C$  is a constant with respect to  $\boldsymbol{\theta}^0$ . This enables the use of MCMC to sample directly from the marginal posterior distribution  $p(\boldsymbol{\theta}^0 | \mathbf{y})$ , and for this purpose we employed the *elliptical slice sampler* of Murray et al. (2010). If the number of data  $\mathbf{y}$  is  $M$ , the computational complexity for marginal likelihood evaluation is  $\mathcal{O}(ND^2 + N^2 D + NM^2 + N^2 M + M^3)$ , which is quadratic with respect to  $N$ . Furthermore, we recognise that the posterior distribution of  $\mathbf{w}^1$  conditional on  $\boldsymbol{\theta}^0$  are Gaussian, meaning that given each sample of  $\boldsymbol{\theta}^0$  and data  $\mathbf{y}$  we can simulate a corresponding sample:

$$\mathbf{w}^1 | \boldsymbol{\theta}^0, \mathbf{y} \sim \mathcal{N}(\mathbf{m}_{**}, \boldsymbol{\Sigma}_{**})$$

where  $\mathbf{m}_{**} := \sigma_\epsilon^{-2} \boldsymbol{\Sigma}_{**} \boldsymbol{\Phi}_0^\top \mathbf{y}$  and  $\boldsymbol{\Sigma}_{**}^{-1} := \boldsymbol{\Sigma}^{-1} + \sigma_\epsilon^{-2} \boldsymbol{\Phi}_0^\top \boldsymbol{\Phi}_0$ . Standard techniques were applied to regularise computation in cases where matrices were poorly conditioned; see Appendix A.4.

This approach produces samples that are asymptotically drawn from the joint posterior  $(\boldsymbol{\theta}^0, \boldsymbol{w}^1) | \mathbf{y}$ . The complexity of this conditional sampling step is  $\mathcal{O}(N^3 + N^2M + NM)$ , which is cubic with respect to  $N$ . This two-stage procedure was observed to work effectively in the experiments that we performed. However, this approach does not extend to deeper BNNs and therefore more general-purpose MCMC or variational inference techniques are likely to be required. The potential for variational techniques to be employed with the ridgelet prior is discussed next.

### 5.3 Variational Inference

Although the algorithm in Section 5.2 was observed to perform well for the experiments reported in Section 6, the limited applicability to shallow networks suggests that an alternative approach to posterior approximation is needed. As we explain in this section, variational inference may be well-suited to this task. In variational inference, a posterior distribution  $p(\boldsymbol{\theta} | \mathbf{y})$  is approximated by a distribution  $q$  selected from a candidate set  $\mathcal{Q}$  via maximisation of an evidence lower bound (ELBO):

$$\max_{q \in \mathcal{Q}} \mathbb{E}_{\boldsymbol{\theta} \sim q} [\log p(\mathbf{y} | \boldsymbol{\theta})] - \text{KL}(q \| \pi)$$

where  $\pi$  is a prior and  $p(\mathbf{y} | \boldsymbol{\theta})$  is the likelihood (Sun et al., 2019). Note that  $\boldsymbol{\theta}$  denotes the set of all parameters of the form  $\boldsymbol{w}_i^l, b_i^l$  involved in the BNN. The term  $\text{KL}(q \| \pi)$  will not be available in closed form in general because of the complicated *a priori* dependency among the parameters in  $\boldsymbol{\theta}$ . However, closer inspection of the ridgelet prior reveals that each layer-wise component of  $\boldsymbol{\theta}$  is conditionally Gaussian given the values of the remaining components; this conditional Gaussian form can be used to circumvent intractability of the Kullback–Leibler term in the ELBO. This suggests a natural approach to optimisation, which cycles through layer-wise subsets of  $\boldsymbol{\theta}$ , maximising the ELBO with respect to each layer-wise component. Although the design and assessment of a variational inference procedure is not a focus of this paper, the possibility to pursue an alternating optimisation procedure demonstrates that there may be room to pursue improved computational methodology for the ridgelet prior if needed.

## 6. Empirical Assessment

In this section we briefly report empirical results that are intended as a proof-of-concept. Our aims in this section are twofold: First, in Section 6.1 we seek to illustrate the theoretical analysis of Section 4 and to explore how well a BNN with a ridgelet prior approximates its intended GP target. Second, in Section 6.2 we aim to establish whether use of the ridgelet prior for a BNN in a Bayesian inferential context confers some of the same characteristics as when the target GP prior is used, in terms of the inferences that are obtained.

Throughout this section we focus on BNNs with a single hidden layer ( $L = 1$ ). The settings for the ridgelet prior were identical to the hyperbolic tangent activation pair  $(\phi, \psi)$  in Table 1 for all experiments reported in the main text; sensitivity to these choices examined in Appendix A.4.3. Our principal interest is in how many hidden units,  $N$ , are required in order for the ridgelet prior to provide an adequate approximation of the associated GP, since the size of the network determines the computational complexity of performing inference

with the ridgelet prior; see Section 5. To this end, we fix  $D$ ,  $\sigma_w$  and  $\sigma_b$  and investigate the influence of  $N$ . The fixed values of the bandwidths  $\sigma_w$  and  $\sigma_b$  were selected by analysing the approximation quality, as described in Appendix A.4.3. For completeness, Table 2 in Appendix A.4.1 reports the values of  $D$ ,  $\sigma_w$  and  $\sigma_b$  that were used for each experiment.

### 6.1 Approximation Quality

The aim of this section is to explore how well a BNN with a ridgelet prior approximates its intended GP target. For this experiment we consider dimension  $d = 1$  and the so-called *squared exponential* covariance function  $k(x, x') := l^2 \exp(-\frac{1}{2s^2}|x - x'|^2)$ , where  $l = 1.0$ ,  $s = 1.5$  were chosen to ensure that samples from the GP demonstrate non-trivial behaviour on our domain of interest  $\mathcal{X} = (-5, 5)$ .

Figure 3, introduced earlier in the paper, presents sample paths from a BNN equipped with the ridgelet prior as the number  $N$  of hidden units is varied from 100 to 3000. In the case  $N = 3000$  the sample paths are almost indistinguishable from those of the target GP. On the other hand,  $N = 3000$  may be a rather large number of units to require, depending on the applied context. Figure 5a explores the approximation quality of the discretised ridgelet transform construction, reporting the *maximum root-mean-square error* (MRMSE)

$$\sup_{\mathbf{x} \in \mathcal{X}} \sqrt{\mathbb{E}[(f(\mathbf{x}) - I_{\sigma, D, N} f(\mathbf{x}))^2 \mid \{\mathbf{w}_i, b_i\}_{i=1}^N]}.$$

It was observed that the approximation error between the BNN and the GP decays at a slow rate, consistent with our theoretical error bound  $O(N^{-\frac{1}{2}})$ . To explore how well the second moments of the GP are being approximated, in Figure 5b we compared the BNN covariance function  $\mathbb{E}[f(x)f(0)]$  as in (9) against the covariance function  $k(x, 0)$  of the target GP. As  $N$  is varied we observe convergence of the BNN covariance function to the GP covariance function. For accurate approximation, a large number of hidden units appears to be required. Additional results, detailing the effect of varying  $D$ ,  $\sigma_w$ ,  $\sigma_b$ , activation function  $\phi$ , and GP covariance  $k$  are provided in Appendix A.4.3.

The MRMSE is a strong assessment that acknowledges the paired nature of  $f$  and  $I_{\sigma, D, N} f$ , in contrast to a more standard comparison of the distributions produced by the BNN and the GP. However, as an alternative criteria to MRMSE, we also computed the maximum mean discrepancy (MMD) between the BNN and the GP; these details and results are reserved for Appendix A.4.2.

### 6.2 Inference and Prediction Using the Ridgelet Prior

In this section we compare the performance of the ridgelet prior to that of its target GP in an inferential context. To this end, we identified tasks where non-trivial prior information is available and can be encoded into a covariance model; the ridgelet prior is then used to approximate this covariance model using a BNN. Three tasks were considered: (i) prediction of atmospheric CO<sub>2</sub> concentration using the well-known Mauna Loa dataset; (ii) prediction of airline passenger numbers using a historical dataset; (iii) a simple in-painting task that is closer in spirit to applications where BNNs may be used. These tasks are toy in their nature and we do not attempt an empirical investigation into the practical value of the ridgelet prior; this will be reserved for a sequel.

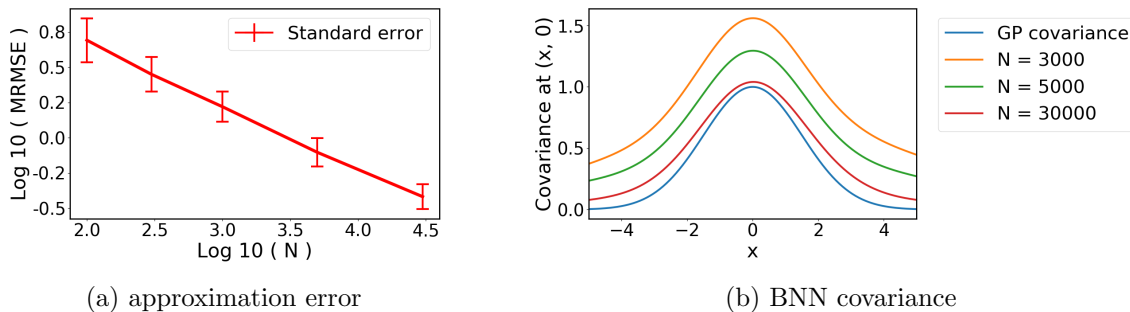


Figure 5: Approximation quality of the ridgelet prior: (a) MRMSE for the BNN approximation of the GP was estimated as the number  $N$  of hidden units was varied with standard errors of 100 independent experiments displayed. (b) The covariance associated to the BNN, as  $N$  is varied, together with the covariance of the GP being approximated.

**Prediction of Atmospheric CO<sub>2</sub>:** In this first task we extracted 43 consecutive months of atmospheric CO<sub>2</sub> concentration data recorded at the Mauna Loa observatory in Hawaii (Keeling and Whorf, 2004) and considered the task of predicting CO<sub>2</sub> concentration up to 23 months ahead. The time period under discussion was re-scaled onto interval  $\mathcal{X} = (-5, 5)$  and the response variable, representing CO<sub>2</sub> concentration, was standardised. This task is well-suited to regression models, such as GPs, that are able to account for non-trivial prior information. In particular, one may seek to encode (1) a linear trend toward increasing concentration of atmospheric CO<sub>2</sub> and (2) a periodic seasonal trend. In the GP framework, this can be achieved using a mean function  $m$  and covariance function  $k$  of the form

$$m(x) := ax, \quad k(x, x') := l^2 \exp\left(-\frac{2}{s^2} \sin\left(\frac{\pi}{p^2}|x - x'|\right)^2\right) \quad (15)$$

where  $a = 0.06$ ,  $l = 1.0$ ,  $s = 0.75$ ,  $p = 1.8$ ; see (Rasmussen and Williams, 2006, Section 5.4.3). The aim of this experiment is to explore whether the posterior predictive distribution obtained using the ridgelet prior for a BNN is similar to that which would be obtained using this GP prior. To this end, we employed a simple Gaussian likelihood

$$y_i = f(x_i, \theta) + \epsilon_i, \quad \epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$$

where  $y_i$  is the standardised CO<sub>2</sub> concentration in month  $i$ ,  $x_i$  is the standardised time corresponding to month  $i$ ,  $x \mapsto f(x, \theta)$  is the regression model and  $\sigma_\epsilon = 0.065$  was fixed. The posterior GP is available in closed form, while we used MCMC to sample from the posterior distribution over the parameters  $\theta$  when a neural network is used, as described in Section 5.2.

Here we present results for BNN with  $N = 500$  hidden units. Our benchmark is the so-called *i.i.d. prior* that takes all parameters to be *a priori* independent with  $w_{i,j}^0 \sim \mathcal{N}(0, \sigma_{w^0})$ ,  $b_i^0 \sim \mathcal{N}(0, \sigma_{b^0})$ , and  $w_{1,i}^1 \sim \mathcal{N}(0, \sigma_{w^1})$ . Here  $\sigma_{w^0} = 3$ ,  $\sigma_{b^0} = 12$  and  $\sigma_{w^1} = 0.1/\sqrt{N}$ . This is to be contrasted with the ridgelet prior, with values of  $\sigma_w = 3$  and  $\sigma_b = 12$  chosen to

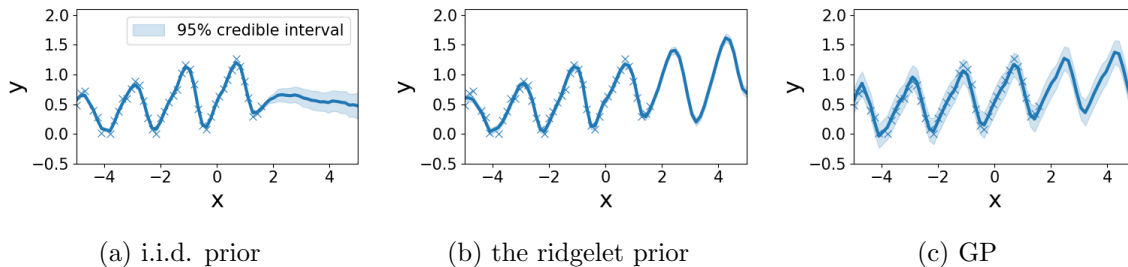


Figure 6: Prediction of Atmospheric CO<sub>2</sub>: Posterior predictive distributions were obtained using (a) a Bayesian neural network (BNN) with an i.i.d. prior, (b) a BNN with the ridgelet prior, and (c) a Gaussian process (GP) prior. The ridgelet prior in (b) was designed to approximate the GP prior used in (c). The solid blue curve is the posterior mean and the shaded regions represent the pointwise 95% credible interval.

ensure a fair comparison with the i.i.d. prior. Figure 6 displays the posterior predictive distributions obtained using (a) the i.i.d. prior, (b) the ridgelet prior, and (c) the original GP. It can be seen that (b) and (c) are more similar than (a) and (c). These results suggest that the ridgelet prior is able to encode non-trivial information on the seasonality of CO<sub>2</sub> concentration into the prior distribution for the parameters of the BNN.

**Prediction of Airline Passenger Numbers:** Next we considered a slightly more challenging example that involves a more intricate periodic trend. The dataset here is a subset of the airline passenger dataset studied in Pearce et al. (2020). The response  $y_i$  represents the (standardised) monthly total number of international airline passengers in the United States and the input  $x_i$  represents the (standardised) time corresponding to month  $i$ . The experimental set-up was identical to the CO<sub>2</sub> experiment, but the prior in (15) was modified to reflect the more intricate periodic trend by taking  $a = 0.2$ ,  $l = 1.0$ ,  $s = 0.75$ ,  $p = 1.75$  and the measurement noise was fixed to  $\sigma_\epsilon = 0.145$ .

In a similar manner to the CO<sub>2</sub> experiment, we implemented the i.i.d. prior and the analogous ridgelet prior, in the latter case based on  $D = 200$  quadrature points. Results in Figure 7 support the conclusion that the ridgelet prior is able to capture this more complicated seasonal trend. In comparison with Pearce et al. (2020) we required more hidden units to perform this regression task. However, we used a standard activation function, where the method of Pearce et al. (2020) would need to develop a new activation function for each covariance model of interest.

**In-Painting Task:** Our final example is a so-called *in-painting* task, where we are required to infer a missing part of an image from the remaining part. Our aim is not to assess the suitability of the ridgelet prior for such tasks – to do so would require extensive and challenging empirical investigation, beyond the scope of the present paper – but rather to validate the ridgelet prior as a proof-of-concept.

The image that we consider is shown in Figure 8a and we censor the central part, described by the red square in Figure 8b. Each pixel corresponds to a real value  $y_i$  and the location of the pixel is denoted  $\mathbf{x}_i \in (-5, 5)^2$ . To the remaining part we add a small amount

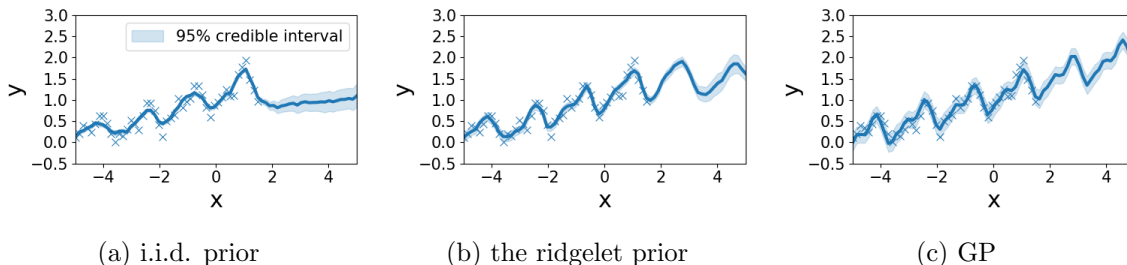


Figure 7: Prediction of Airline Passenger Numbers: Posterior predictive distributions were obtained using (a) a Bayesian neural network (BNN) with an i.i.d. prior, (b) a BNN with the ridgelet prior, and (c) a Gaussian process (GP) prior. The ridgelet prior in (b) was designed to approximate the GP prior used in (c). The solid blue curve is the posterior mean and the shaded regions represent the pointwise 95% credible interval.

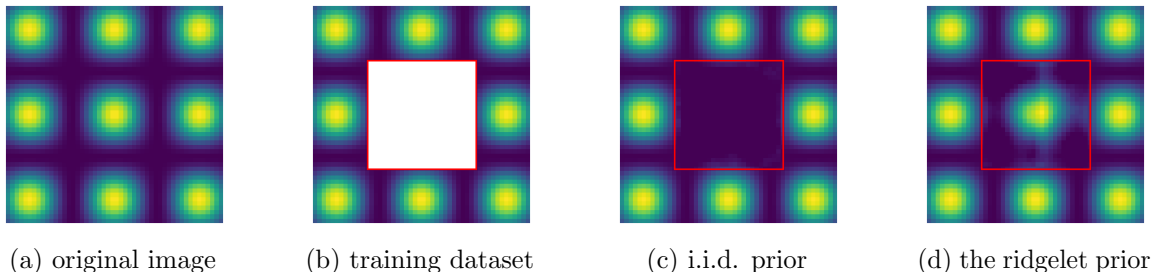


Figure 8: In-Painting Task: The central part of the original image in (a) was censored to produce (b) and the in-painting task is to infer the missing part of the image using the remaining part as a training dataset. Posterior predictive distributions were obtained using (c) a Bayesian neural network (BNN) with an i.i.d. prior, and (d) a BNN with the ridgelet prior.

of i.i.d. noise,  $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$  to each pixel  $i$ , in Figure 8b (the addition of noise here ensures conditions for ergodicity of MCMC are satisfied; otherwise the posterior is supported on a submanifold of the parameter space and more sophisticated sampling procedures would be required). The task is then to infer this missing central region using the remaining part of the image as a training dataset.

For the statistical regression model we considered a GP whose mean function  $m$  is zero and covariance function  $k$  is

$$k(\mathbf{x}, \mathbf{x}') := l^2 \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2s^2}\right) + \left(\cos\left(\frac{\pi x_1}{2}\right) + 1\right)\left(\cos\left(\frac{\pi x_2}{2}\right) + 1\right)\left(\cos\left(\frac{\pi x'_1}{2}\right) + 1\right)\left(\cos\left(\frac{\pi x'_2}{2}\right) + 1\right), \tag{16}$$

where  $l = 0.1$ ,  $s = 0, 1$ . The periodic structure induced by cosine functions is deliberately chosen to be commensurate with the separation between modes in the original dataset,

meaning that considerable prior information is provided in the GP covariance model. Our benchmark here was a BNN equipped with  $N = 2000$  hidden units and an i.i.d. prior, constructed in an analogous manner as before with parameters  $\sigma_{w^0} = 2$ ,  $\sigma_{b^0} = 12$  and  $\sigma_{w^1} = 0.1/\sqrt{N}$ . The ridgelet prior was based on a BNN of the same size, with  $\sigma_w = 2$  and  $\sigma_b = 12$  to ensure a fair comparison with the i.i.d. prior.

Figure 8c and Figure 8d show posterior means estimates for the missing region in the in-painting task, respectively for the i.i.d. prior and the ridgelet prior. It is interesting to observe that the i.i.d. prior gives rise to a posterior mean that is approximately constant, whereas the ridgelet prior gives rise to a posterior mean that somewhat resembles the original image. This suggests that the specific structure of the covariance model in (16) has been faithfully encoded into the ridgelet prior, leading to superior performance on this stylised in-painting task.

### 6.3 Limitations of the Ridgelet Prior for Deep Networks

Finally, in this section we examine the performance of the ridgelet prior in a setting where the depth of the network is increased. For this experiment, we fixed the hidden unit number  $N_l$  of each hidden layer  $l$  to  $N_l = 10,000$  and increased the number of hidden layers from  $L = 1$  to  $L = 5$ . The approximation error between a BNN with the ridgelet prior and the target GP was quantified using MMD, aforementioned in Section 6.1: see Appendix A.4.2 for detail. Results are displayed in Figure 9. Interestingly, the lowest value of MMD was observed at  $L = 1$ , suggesting that the performance of the ridgelet prior deteriorates when the depth of the network is increased. This makes intuitive sense, since with each additional layer in the network an additional discretisation of the ridgelet transform is required. This discretisation error then has an opportunity to accumulate and propagate through the network. It would be interesting to explore strategies to mitigate this degradation of performance, but our focus in this work was limited to providing theoretical analysis and an empirical proof-of-concept.

## 7. Conclusion

One of the main barriers to the wide-spread adoption of BNN is the identification of prior distributions that are meaningful when lifted to the output space of the network. In this paper it was shown that the ridgelet transform facilitates the consistent approximation of a GP using a BNN. This has the potential to bring the powerful framework of covariance modelling for GPs to bear on the task of prior specification for BNN. In contrast to earlier work in this direction (Flam-Shepherd et al., 2017; Hafner et al., 2020; Pearce et al., 2020; Sun et al., 2019), our construction is accompanied by theoretical analysis that establishes the approximation is consistent. Moreover, we are able to provide a finite-sample-size error bound that requires only weak assumptions on the GP covariance model (i.e. that the mean and covariance function is continuous and differentiable).

This role of this paper was to establish the ridgelet prior as a theoretical proof-of-concept only and there remain several open questions to be addressed:

- In real applications, is it necessary to have an accurate approximation of the intended covariance model, in order to deliver improved performance of the BNN, or is a crude approximation sufficient?

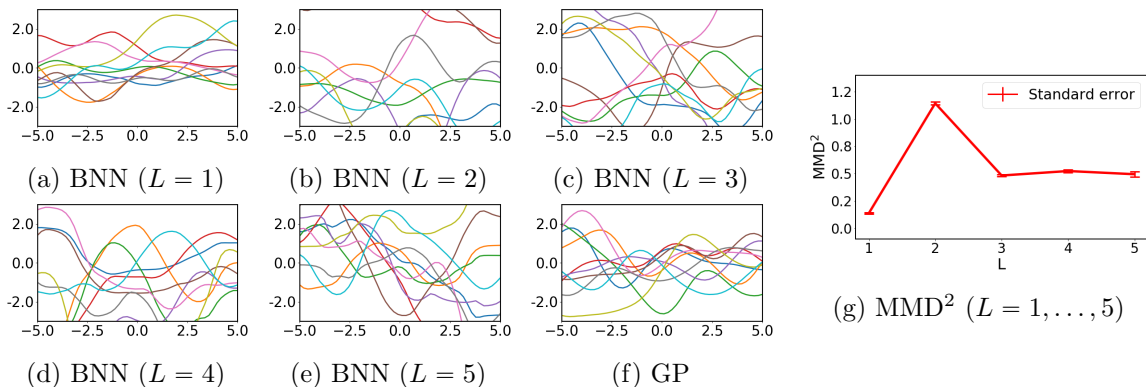


Figure 9: Sample paths from Bayesian neural networks (BNNs) equipped with the ridgelet prior as the number of hidden layers  $L$  is increased. The number of hidden units was fixed to  $N_l = 10,000$  for each hidden layer  $l = 1, \dots, L$ . In (f) sample paths are shown from the target Gaussian process (GP) prior. In (g) the  $MMD^2$  between the BNN and the GP are displayed. The standard error is calculated by 10 independent computations.

- What cubature rules  $\{\mathbf{x}_i, u_i\}_{i=1}^D$  are most effective in applications - is a regular grid needed, or do real-world applications exhibit an effective low dimension so that e.g. a sparse grid could be used?
- In the case of multiple hidden layers, can the convergence of the ridgelet prior to a *deep GP* (Damianou and Lawrence, 2013; Dunlop et al., 2018) be established?

These questions will likely require a substantial amount of work to address in full, but we hope to pursue some of them in a sequel.

**Acknowledgements** The authors were supported by the Lloyd’s Register Foundation Programme on Data-Centric Engineering and the Alan Turing Institute under the EPSRC grant [EP/N510129/1].



## Appendix A. Supplementary Material

This appendix is structured as follows:

A.1	Convergence of the Finite-Bandwidth Ridgelet Transform . . . . .	25
A.1.1	Approximate Identity Operators . . . . .	27
A.1.2	Proof of Theorem 2 . . . . .	28
A.1.3	Auxiliary Lemma 1 . . . . .	33
A.2	Proof of Theorem 1 . . . . .	34
A.2.1	Technical Lemmas . . . . .	34
A.2.2	An Intermediate Ridgelet Reconstruction Result . . . . .	37
A.2.3	Proof of Theorem 1 . . . . .	41
A.3	An Analogous Result to Theorem 1 for Unbounded $\phi$ . . . . .	44
A.4	Experiments . . . . .	48
A.4.1	Experimental Setting . . . . .	48
A.4.2	Alternative Error Measurement by MMD . . . . .	49
A.4.3	Comparison of Prior Predictive with Different Settings . . . . .	49

*Notation:* Throughout this appendix we adopt identical notation to the main text, but for brevity we denote  $\|\cdot\|_{L^1(\mathcal{X})}$  as  $\|\cdot\|_{L^1}$  and  $\|\cdot\|_{L^\infty(\mathcal{X})}$  as  $\|\cdot\|_{L^\infty}$  whenever  $\mathcal{X}$  is the Euclidean space  $\mathbb{R}^d$  of any dimension  $d \in \mathbb{N}$ . We also introduce function classes of higher order differentiability. Denote by  $C^r(\mathcal{X})$  the set of all functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  for which the derivatives  $\partial^\alpha f$  exist and are continuous on  $\mathcal{X}$  for all  $\alpha \in \mathbb{N}_0^d$  s.t.  $|\alpha| \leq r$ . Denote by  $C^{r \times r}(\mathcal{X} \times \mathcal{X})$  the set of all functions  $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  for which the derivatives  $\partial^{\alpha, \beta} h$  exist and are continuous on  $\mathcal{X} \times \mathcal{X}$  for all  $\alpha, \beta \in \mathbb{N}_0^d$  s.t.  $|\alpha|, |\beta| \leq r$ .

### A.1 Convergence of the Finite-Bandwidth Ridgelet Transform

In this section we derive a finite-sample-size error bound for the finite-bandwidth ridgelet transform of Definition 2. Our analysis is more general than the Gaussian case presented in the main text, and we consider more general probability measures on  $\mathbf{w}$  and  $b$  to control convergence of  $\lambda_\sigma$  toward the improper uniform limit.

Let  $p_{\mathbf{w}}$  and  $p_b$  be probability densities, respectively, on  $\mathbb{R}^d$  and  $\mathbb{R}$ , satisfying regularity requirements that will shortly be described. For  $0 < \sigma_{\mathbf{w}} < \infty, 0 < \sigma_b < \infty$ , define scaled densities of  $p_{\mathbf{w}}$  and  $p_b$  by  $p_{\mathbf{w}, \sigma}(\mathbf{w}) := \sigma_{\mathbf{w}}^{-d} p_{\mathbf{w}}(\sigma_{\mathbf{w}}^{-1} \mathbf{w})$  and  $p_{b, \sigma}(b) := \sigma_b^{-1} p_b(\sigma_b^{-1} b)$ . The parameters  $\sigma_{\mathbf{w}}$  and  $\sigma_b$  will be called *bandwidths*. To recover the case presented in the main text, if  $p_{\mathbf{w}}$  and  $p_b$  are standard Gaussians then  $p_{\mathbf{w}, \sigma}(\mathbf{w})$  and  $p_{b, \sigma}(b)$  are Gaussians with variances  $\sigma_{\mathbf{w}}^2 \mathbf{I}$  and  $\sigma_b^2$ , i.e.

$$p_{\mathbf{w}, \sigma}(\mathbf{w}) = \frac{1}{(2\pi\sigma_{\mathbf{w}}^2)^{d/2}} \exp\left(-\frac{\|\mathbf{w}\|^2}{2\sigma_{\mathbf{w}}^2}\right), \quad p_{b, \sigma}(b) = \frac{1}{(2\pi\sigma_b^2)^{1/2}} \exp\left(-\frac{b^2}{2\sigma_b^2}\right).$$

Then let  $\lambda_\sigma(\mathbf{w}, b)$  be a measure whose density is  $Z p_{\mathbf{w}, \sigma}(\mathbf{w}) p_{b, \sigma}(b)$ , where for the analysis that follows it will be convenient to set

$$Z := \frac{(2\pi)^{\frac{1}{2}} \sigma_{\mathbf{w}}^d \sigma_b}{\|\widehat{p}_{\mathbf{w}}\|_{L^1(\mathbb{R}^d)} \|\widehat{p}_b\|_{L^1(\mathbb{R})}},$$

indicating the total measure assigned by  $\lambda_\sigma$  to  $\mathbb{R}^{d+1}$ . For the case of standard Gaussian densities  $p_{\mathbf{w}}$  and  $p_b$  presented in the main text, we have  $\|\widehat{p}_{\mathbf{w}}\|_{L^1(\mathbb{R}^d)} = \|\widehat{p}_b\|_{L^1(\mathbb{R})} = 1$ .

Our analysis covers general  $p_{\mathbf{w}}$  and  $p_b$  provided that certain regularity conditions are satisfied. These will now be described. First, consider an arbitrary probability density function  $p$  defined on  $\mathbb{R}^m$  for some  $m \in \mathbb{N}$  and consider the following properties that  $p$  may satisfy:

- (i) (Symmetry at the origin)  $p(\mathbf{x}) = p(-\mathbf{x}), \forall \mathbf{x} \in \mathbb{R}^m$ ,
- (ii) (Boundedness)  $\|p\|_{L^\infty(\mathbb{R}^m)} < \infty$ ,
- (iii) (Finite second moment)  $\int_{\mathbb{R}^m} \|\mathbf{x}\|^2 p(\mathbf{x}) d\mathbf{x} < \infty$ ,
- (iv) (Positivity of the Fourier transform)  $\widehat{p}(\mathbf{x}) > 0, \forall \mathbf{x} \in \mathbb{R}^m$ ,
- (v) (Integrable Fourier transform)  $\|\widehat{p}\|_{L^1(\mathbb{R}^m)} < \infty$ .

Of these properties, only (iv) and (v) are not straight forward. If  $\widehat{p}$  can be computed then conditions (iv) and (v) can be directly verified. Otherwise, sufficient conditions on  $p$  for (iv) can be found in Tuck (2006) and sufficient conditions on  $p$  for (v) can be found in Lifyand (2016). The regularity that we require on  $p_{\mathbf{w}}$  and  $p_b$  can now be specified:

**Assumption 4** (Finite bandwidth). *The probability density functions  $p_{\mathbf{w}} : \mathbb{R}^d \rightarrow [0, \infty)$  and  $p_b : \mathbb{R} \rightarrow [0, \infty)$  satisfy properties (i) – (v) above and, in addition,*

1.  $\int_{\mathbb{R}^d} \|\mathbf{x}\|^2 \widehat{p}_{\mathbf{w}}(\mathbf{x}) d\mathbf{x} < \infty$ ,
2.  $\|\partial p_b\|_{L^\infty(\mathbb{R})} < \infty$ ,

where we recall that  $\partial p_b$  denotes the first derivative of  $p_b$ .

Now we present the convergence result of interest. Let  $M_1(f) := \max_{|\alpha| \leq 1} \sup_{\mathbf{x} \in \mathbb{R}^d} |\partial^\alpha f(\mathbf{x})|$  and  $B_1(f) := \int_{\mathbb{R}^d} |f(\mathbf{x})| (1 + \|\mathbf{x}\|_2) d\mathbf{x}$ .

**Theorem 2.** *Let Assumption 1 and Assumption 4 hold, and let  $f \in C^1(\mathbb{R}^d)$  satisfy  $M_1(f) < \infty$  and  $B_1(f) < \infty$ . Then*

$$\sup_{\mathbf{x} \in \mathbb{R}^d} |f(\mathbf{x}) - (R_\sigma^* R)[f](\mathbf{x})| \leq C \max(M_1(f), B_1(f)) \left\{ \frac{1}{\sigma_{\mathbf{w}}} + \frac{\sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)}{\sigma_b} \right\}$$

for some constant  $C$  that is independent of  $\sigma_{\mathbf{w}}, \sigma_b$  and  $f$ , but may depend on  $\phi, \psi, p_{\mathbf{w}}$  and  $p_b$ .

Our priority here was to provide a simple upper bound on the reconstruction error, that separates the influence of  $f$  from the influence of the bandwidths  $\sigma_{\mathbf{w}}$  and  $\sigma_b$ ; the bound is not claimed to be tight. The formulation of Theorem 2 enables us to conveniently conclude that taking  $\sigma_{\mathbf{w}}$  and  $\sigma_b$  to infinity in a manner such that  $\sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1) / \sigma_b \rightarrow 0$  will lead to consistent recovery of  $f$ , for any function  $f$  for which  $M_1(f), B_1(f) < \infty$  is satisfied. The term  $M_1(f)$  reflects the general difficulty of approximating  $f$  using a finite-bandwidth ridgelet transform, while the term  $B_1(f)$  serves to deal with the tails of  $f$  and the fact that a supremum is

taken over the unbounded domain  $\mathbb{R}^d$ . Theorem 5 in Appendix A.3 presents a variation on Theorem 2 that relaxes the boundedness assumption on the activation function from  $\phi \in C_0^*(\mathbb{R})$  to  $\phi \in C_1^*(\mathbb{R})$  at the expense of changing the tail condition from  $B_1(f) < \infty$  to  $B_2(f) := \int_{\mathbb{R}^d} |f(\mathbf{x})|(1 + \|\mathbf{x}\|_2)^2 d\mathbf{x} < \infty$  and restricting the supremum to a bounded subset of  $\mathbb{R}^d$ .

The proof of Theorem 2 makes use of techniques from the theory of *approximate identity* integral operators, which we recall next in Appendix A.1.1. The proof of Theorem 2 is then presented in Appendix A.1.2. The proof relies on an auxiliary technical lemma regarding the Fourier transform, whose statement and proof we defer to Appendix A.1.3.

### A.1.1 APPROXIMATE IDENTITY OPERATORS

An *approximate identity* operator is an integral operator which converges to an identity operator in an appropriate limit. The discussion in this section follows Ginè and Nickl (Ginè and Nickl, 2015, ch. 4.1.3, 4.3.6). For  $h > 0$ , define an operator  $K_h$  by

$$K_h[f](\mathbf{x}) := \int_{\mathbb{R}^d} f(\mathbf{x}') \frac{1}{h^d} K\left(\frac{\mathbf{x}}{h}, \frac{\mathbf{x}'}{h}\right) d\mathbf{x}' \quad (17)$$

where  $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  is a measurable function and  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is suitably regular for the integral to exist and be well-defined. Proposition 1 provides sufficient conditions for the approximate identity  $K_h[f]$  to converge to an identity operator when  $h \rightarrow 0$ .

**Proposition 1.** (Ginè and Nickl, 2015, Propositions 4.3.31 and 4.3.33, p.368) *Let  $K_h[f]$  be defined as in (17) with  $K$  a measurable function satisfying, for some  $N \in \mathbb{N}_0$ ,*

1.  $\int_{\mathbb{R}^d} \sup_{\mathbf{v} \in \mathbb{R}^d} |K(\mathbf{v}, \mathbf{v} - \mathbf{u})| \|\mathbf{u}\|^N d\mathbf{u} < \infty$ ,
2. for all  $\mathbf{v} \in \mathbb{R}^d$  and all multi-indices  $\boldsymbol{\alpha}$  s.t.  $|\boldsymbol{\alpha}| \in \{1, \dots, N - 1\}$ ,
  - (a)  $\int_{\mathbb{R}^d} K(\mathbf{v}, \mathbf{v} - \mathbf{u}) d\mathbf{u} = 1$ ,
  - (b)  $\int_{\mathbb{R}^d} K(\mathbf{v}, \mathbf{v} - \mathbf{u}) \mathbf{u}^\alpha d\mathbf{u} = 0$ .

Then for each  $m \leq N$  there exists a constant  $C$ , depending only on  $m$  and  $K$ , such that

$$f \in C^m(\mathbb{R}^d) \implies \sup_{\mathbf{x} \in \mathbb{R}^d} |K_h[f](\mathbf{x}) - f(\mathbf{x})| \leq Ch^m \max_{|\boldsymbol{\alpha}|=m} \sup_{\mathbf{x} \in \mathbb{R}^d} |\partial^\alpha f(\mathbf{x})|.$$

The sense in which Proposition 1 will be used in the proof of Theorem 2 is captured by the following example:

**Example 3.** *Consider a translation invariant kernel of the form  $K(\mathbf{x}, \mathbf{x}') = \varphi(\mathbf{x} - \mathbf{x}')$  for some  $\varphi : \mathbb{R}^d \rightarrow (0, \infty)$ . Further assume  $\int_{\mathbb{R}^d} \varphi(\mathbf{u}) d\mathbf{u} = 1$  and  $\int_{\mathbb{R}^d} \varphi(\mathbf{u}) \|\mathbf{u}\|^2 d\mathbf{u} < \infty$ . Then  $K$  satisfies the preconditions of Proposition 1 for  $N = 2$  and hence  $K_h[f]$  is an approximate identity operator.*

## A.1.2 PROOF OF THEOREM 2

For this section it is convenient to introduce a notational shorthand. Let  $\epsilon_{\mathbf{w}} := \sigma_{\mathbf{w}}^{-1}$  and  $\epsilon_b := \sigma_b^{-1}$ , so that  $p_{\mathbf{w}}(\sigma_{\mathbf{w}}^{-1}\mathbf{w}) = p_{\mathbf{w}}(\epsilon_{\mathbf{w}}\mathbf{w})$  and  $p_b(\sigma_b^{-1}b) = p_b(\epsilon_b b)$ . Further introduce the shorthand  $p_{\mathbf{w}}^\epsilon(\cdot) := p_{\mathbf{w}}(\epsilon_{\mathbf{w}}\cdot)$  and  $p_b^\epsilon(\cdot) := p_b(\epsilon_b\cdot)$ .

Now we turn to the proof of Theorem 2. Denote the topological dual space of  $\mathcal{S}(\mathbb{R}^d)$  by  $\mathcal{S}'(\mathbb{R}^d)$ . The elements of  $\mathcal{S}'(\mathbb{R}^d)$  are called *tempered distributions* and the generalized Fourier transform can be considered as the Fourier transform on tempered distributions  $\mathcal{S}'(\mathbb{R}^d)$  (Grafakos, 2000, p.123-131). Throughout the proof we exchange the order of integrals; we do so only when the absolute value of the integrand is itself an integrable function, so that from Fubini's theorem the interchange can be justified.

**Proof** [Proof of Theorem 2] The goal is to bound reconstruction error when the Lebesgue measure in the classical ridgelet transform is replaced by the finite measure

$$d\lambda_\sigma(\mathbf{w}, b) = (2\pi)^{\frac{1}{2}} \|\widehat{p}_{\mathbf{w}}\|_{L^1}^{-1} \|\widehat{p}_b\|_{L^1}^{-1} p_{\mathbf{w}}(\sigma_{\mathbf{w}}^{-1}\mathbf{w}) p_b(\sigma_b^{-1}b) d\mathbf{w} db.$$

The reconstruction  $(R_\sigma^* R)[f]$  of a function  $f$  on  $\mathbb{R}^d$  is defined by the following integral:

$$\begin{aligned} & (R_\sigma^* R)[f](\mathbf{x}) \\ &= \int_{\mathbb{R}^{d+1}} \int_{\mathbb{R}^d} f(\mathbf{x}') \psi(\mathbf{w} \cdot \mathbf{x}' + b) d\mathbf{x}' \phi(\mathbf{w} \cdot \mathbf{x} + b) d\lambda_\sigma(\mathbf{w}, b) \\ &= \int_{\mathbb{R}^d} f(\mathbf{x}') \left\{ (2\pi)^{\frac{1}{2}} \|\widehat{p}_{\mathbf{w}}\|_{L^1}^{-1} \|\widehat{p}_b\|_{L^1}^{-1} \int_{\mathbb{R}} \int_{\mathbb{R}^d} \psi(\mathbf{w} \cdot \mathbf{x}' + b) \phi(\mathbf{w} \cdot \mathbf{x} + b) p_{\mathbf{w}}^\epsilon(\mathbf{w}) p_b^\epsilon(b) d\mathbf{w} db \right\} d\mathbf{x}'. \end{aligned}$$

It is thus clear that  $(R_\sigma^* R)[f]$  is in some sense a smoothed version of  $f$ , and in what follows we will make use of the theory of approximate identity operators discussed in Appendix A.1.1. We aim to deal separately with the reconstruction error due to the use of a finite measure on  $\mathbf{w}$  and the reconstruction error due to the use of a finite measure on  $b$ . To this end, it will be convenient to (formally) define a new operator  $R_\sigma^* R^{(w)}$  by

$$(R_\sigma^* R^{(w)})[f](\mathbf{x}) := \int_{\mathbb{R}^d} f(\mathbf{x}') \left( \|\widehat{p}_{\mathbf{w}}\|_{L^1}^{-1} \int_{\mathbb{R}} \int_{\mathbb{R}^d} \psi(\mathbf{w} \cdot \mathbf{x}' + b) \phi(\mathbf{w} \cdot \mathbf{x} + b) p_{\mathbf{w}}^\epsilon(\mathbf{w}) d\mathbf{w} db \right) d\mathbf{x}'$$

which replaces  $p_b^\epsilon$  by the Lebesgue measure on  $\mathbb{R}$ . That is,  $(R_\sigma^* R^{(w)})[f]$  can be intuitively considered as an idealised version of  $(R_\sigma^* R)[f]$  where the reconstruction error due to the use of a finite measure on  $b$  is removed. Our analysis will then proceed based on the following triangle inequality:

$$\begin{aligned} & \sup_{\mathbf{x} \in \mathbb{R}^d} |f(\mathbf{x}) - (R_\sigma^* R)[f](\mathbf{x})| \\ & \leq \underbrace{\sup_{\mathbf{x} \in \mathbb{R}^d} |f(\mathbf{x}) - (R_\sigma^* R^{(w)})[f](\mathbf{x})|}_{(*)} + \underbrace{\sup_{\mathbf{x} \in \mathbb{R}^d} |(R_\sigma^* R^{(w)})[f](\mathbf{x}) - (R_\sigma^* R)[f](\mathbf{x})|}_{(**)}. \quad (18) \end{aligned}$$

Different strategies are required to bound  $(*)$  and  $(**)$  and we address them separately next.

**Bounding  $(*)$ :** A bound on  $(*)$  uses techniques from approximate identity operators described in Appendix A.1.1. To this end, we show that  $R_\sigma^* R$  satisfies the preconditions of Proposition 1 to obtain a bound.

Define  $\tilde{k} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  by

$$\tilde{k}(\mathbf{x}, \mathbf{x}') := \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \int_{\mathbb{R}} \int_{\mathbb{R}^d} \phi(\mathbf{w} \cdot \mathbf{x} + b) \psi(\mathbf{w} \cdot \mathbf{x}' + b) p_{\mathbf{w}}(\mathbf{w}) d\mathbf{w} db. \quad (19)$$

From a change of variable  $b' = \mathbf{w} \cdot \mathbf{x} + b$ ,

$$\tilde{k}(\mathbf{x}, \mathbf{x}') = \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \int_{\mathbb{R}^d} \int_{\mathbb{R}} \phi(b') \psi(\mathbf{w} \cdot (\mathbf{x}' - \mathbf{x}) + b') db' p_{\mathbf{w}}(\mathbf{w}) d\mathbf{w}. \quad (20)$$

Next we perform some Fourier analysis similar to that in (Sonoda and Murata, 2017, Appendix C). From the discussion on generalised Fourier transform in Section 4.1, recall that  $\hat{t} \in L^2_{\text{loc}}(\mathbb{R}^d \setminus \{\mathbf{0}\})$  is a *generalised Fourier transform* of function  $t$  if there exists an integer  $m \in \mathbb{N}_0$  such that  $\int_{\mathbb{R}^d} \hat{t}(\mathbf{w}) h(\mathbf{w}) d\mathbf{w} = \int_{\mathbb{R}^d} t(\mathbf{x}) \hat{h}(\mathbf{x}) d\mathbf{x}$  for all  $h \in \mathcal{S}_{2m}(\mathbb{R}^d)$ . We set  $t = \phi$  and  $\hat{h} = \psi$  for our analysis. Part (3) of Assumption 1 implies that  $\psi \in \mathcal{S}_{2m}(\mathbb{R}^d)$  for some  $m$  depending on the generalised Fourier transform  $\hat{\phi}$ . Let  $\psi_-(b) := \psi(-b)$ . From the definition of the generalised Fourier transform,

$$\begin{aligned} \int_{\mathbb{R}} \phi(b') \psi(\mathbf{w} \cdot (\mathbf{x}' - \mathbf{x}) + b') db' &= \int_{\mathbb{R}} \phi(b') \psi_-(-b' - \mathbf{w} \cdot (\mathbf{x}' - \mathbf{x})) db' \\ &= \int_{\mathbb{R}} \hat{\phi}(\xi_1) \overline{\widehat{\psi}(\xi_1)} e^{i\xi_1 \mathbf{w} \cdot (\mathbf{x} - \mathbf{x}')} d\xi_1, \end{aligned} \quad (21)$$

where we used the fact that  $\widehat{\psi_-} = \overline{\widehat{\psi}}$  for the real function  $\psi_-(b)$  for the last equality (Grafakos, 2000, p.109,113).

Note that  $\xi_1 \mapsto \hat{\phi}(\xi_1) \overline{\widehat{\psi}(\xi_1)}$  belongs to  $L^1(\mathbb{R})$  from part (3) of Assumption 1 and therefore (21) exists. Also we have that

$$\widehat{p_{\mathbf{w}}}(\xi_1 \mathbf{x}' - \xi_1 \mathbf{x}) = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} e^{i\xi_1 \mathbf{w} \cdot (\mathbf{x} - \mathbf{x}')} p_{\mathbf{w}}(\mathbf{w}) d\mathbf{w}. \quad (22)$$

Substituting (21) and (22) into (20) gives that

$$\begin{aligned} \tilde{k}(\mathbf{x}, \mathbf{x}') &= \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \int_{\mathbb{R}^d} \int_{\mathbb{R}} \phi(b') \psi(\mathbf{w} \cdot (\mathbf{x}' - \mathbf{x}) + b') db' p_{\mathbf{w}}(\mathbf{w}) d\mathbf{w} \\ &= \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \int_{\mathbb{R}} \hat{\phi}(\xi_1) \overline{\widehat{\psi}(\xi_1)} \int_{\mathbb{R}^d} e^{i\xi_1 \mathbf{w} \cdot (\mathbf{x} - \mathbf{x}')} p_{\mathbf{w}}(\mathbf{w}) d\mathbf{w} d\xi_1 \\ &= (2\pi)^{\frac{d}{2}} \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \int_{\mathbb{R}} \hat{\phi}(\xi_1) \overline{\widehat{\psi}(\xi_1)} \widehat{p_{\mathbf{w}}}(\xi_1 \mathbf{x}' - \xi_1 \mathbf{x}) d\xi_1. \end{aligned}$$

Let  $k^{\mathbf{w}}(\mathbf{v}) := \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \widehat{p_{\mathbf{w}}}(\mathbf{v})$ , so that

$$\tilde{k}(\mathbf{x}, \mathbf{x}') = (2\pi)^{\frac{d}{2}} \int_{\mathbb{R}} \hat{\phi}(\xi) \overline{\widehat{\psi}(\xi)} k^{\mathbf{w}}(\xi \mathbf{x} - \xi \mathbf{x}') d\xi. \quad (23)$$

Recall that  $p_{\mathbf{w}}^{\epsilon}(\mathbf{w}) = p_{\mathbf{w}}(\epsilon \mathbf{w})$  and let  $k_{\epsilon}^{\mathbf{w}}(\mathbf{v}) := \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \widehat{p_{\mathbf{w}}^{\epsilon}}(\mathbf{v})$ . Since the Fourier transform of  $\mathbf{w} \mapsto p_{\mathbf{w}}(\epsilon \mathbf{w})$  is to be  $\mathbf{v} \mapsto \frac{1}{\epsilon^d} \widehat{p_{\mathbf{w}}}(\frac{\mathbf{v}}{\epsilon \mathbf{w}})$  by standard properties of the Fourier transform (Grafakos, 2000, p.109,113),  $k_{\epsilon}^{\mathbf{w}}(\mathbf{v}) = \frac{1}{\epsilon^d} k^{\mathbf{w}}(\frac{\mathbf{v}}{\epsilon \mathbf{w}})$ . Define

$$\tilde{k}_{\epsilon}(\mathbf{x}, \mathbf{x}') := \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \int_{\mathbb{R}} \int_{\mathbb{R}^d} \phi(\mathbf{w} \cdot \mathbf{x} + b) \psi(\mathbf{w} \cdot \mathbf{x}' + b) p_{\mathbf{w}}^{\epsilon}(\mathbf{w}) d\mathbf{w} db. \quad (24)$$

Noting the similarity between (24) and (19), an analogous argument to that just presented shows that we may re-express (24) as

$$\tilde{k}_\epsilon(\mathbf{x}, \mathbf{x}') = (2\pi)^{\frac{d}{2}} \int_{\mathbb{R}} \widehat{\phi}(\xi) \overline{\widehat{\psi}(\xi)} k_\epsilon^{\mathbf{w}}(\xi \mathbf{x} - \xi \mathbf{x}') d\xi = \frac{1}{\epsilon_{\mathbf{w}}^d} \tilde{k}\left(\frac{\mathbf{x}}{\epsilon_{\mathbf{w}}}, \frac{\mathbf{x}'}{\epsilon_{\mathbf{w}}}\right).$$

Then

$$R_\sigma^* R^{(w)} f(\mathbf{x}) = \int_{\mathbb{R}^d} f(\mathbf{x}') \tilde{k}_\epsilon(\mathbf{x}, \mathbf{x}') d\mathbf{x}' = \int_{\mathbb{R}^d} f(\mathbf{x}') \frac{1}{\epsilon_{\mathbf{w}}^d} \tilde{k}\left(\frac{\mathbf{x}}{\epsilon_{\mathbf{w}}}, \frac{\mathbf{x}'}{\epsilon_{\mathbf{w}}}\right) d\mathbf{x}'.$$

Now we will show that  $\tilde{k}$  satisfies the pre-conditions of Proposition 1. That is, setting  $N = 2$ , we will show that: For all  $\mathbf{v} \in \mathbb{R}^d$  and all multi-index  $\boldsymbol{\alpha} \in \mathbb{N}_0^d$  s.t.  $|\boldsymbol{\alpha}| = 1$ ,

$$\int_{\mathbb{R}^d} \sup_{\mathbf{v} \in \mathbb{R}^d} |\tilde{k}(\mathbf{v}, \mathbf{v} - \mathbf{u})| \|\mathbf{u}\|^2 d\mathbf{u} < \infty, \quad (25)$$

$$\int_{\mathbb{R}^d} \tilde{k}(\mathbf{v}, \mathbf{v} - \mathbf{u}) d\mathbf{u} = 1, \quad (26)$$

$$\int_{\mathbb{R}^d} \tilde{k}(\mathbf{v}, \mathbf{v} - \mathbf{u}) \mathbf{u}^\alpha d\mathbf{u} = 0. \quad (27)$$

First we verify (25). From (23) we have that

$$\begin{aligned} \int_{\mathbb{R}} \sup_{\mathbf{v} \in \mathbb{R}^d} |\tilde{k}(\mathbf{v}, \mathbf{v} - \mathbf{u})| \|\mathbf{u}\|^2 d\mathbf{u} &= (2\pi)^{\frac{d}{2}} \int_{\mathbb{R}^d} \left| \int_{\mathbb{R}} \overline{\widehat{\psi}(\xi)} \widehat{\phi}(\xi_1) k^{\mathbf{w}}(\xi \mathbf{u}) d\xi \right| \|\mathbf{u}\|^2 d\mathbf{u} \\ &\leq (2\pi)^{\frac{d}{2}} \int_{\mathbb{R}} \int_{\mathbb{R}^d} |\overline{\widehat{\psi}(\xi)} \widehat{\phi}(\xi)| |k^{\mathbf{w}}(\xi \mathbf{u})| \|\mathbf{u}\|^2 d\mathbf{u} d\xi. \end{aligned}$$

By the change of variables  $\mathbf{u}' = \xi \mathbf{u}$ ,

$$\begin{aligned} \int_{\mathbb{R}} \sup_{\mathbf{v} \in \mathbb{R}^d} |\tilde{k}(\mathbf{v}, \mathbf{v} - \mathbf{u})| \|\mathbf{u}\|^2 d\mathbf{u} &\leq (2\pi)^{\frac{d}{2}} \int_{\mathbb{R}^d} \int_{\mathbb{R}} |\overline{\widehat{\psi}(\xi)} \widehat{\phi}(\xi)| |k^{\mathbf{w}}(\mathbf{u}')| \left\| \frac{\mathbf{u}'}{\xi} \right\|^2 \frac{1}{|\xi|^d} d\mathbf{u}' d\xi \\ &\leq (2\pi)^{\frac{d}{2}} \int_{\mathbb{R}} \int_{\mathbb{R}^d} \frac{|\overline{\widehat{\psi}(\xi)} \widehat{\phi}(\xi)|}{|\xi|^{d+2}} |k^{\mathbf{w}}(\mathbf{u}')| \|\mathbf{u}'\|^2 d\mathbf{u}' d\xi \\ &\leq (2\pi)^{\frac{d}{2}} \int_{\mathbb{R}} \frac{|\overline{\widehat{\psi}(\xi)} \widehat{\phi}(\xi)|}{|\xi|^{d+2}} d\xi \int_{\mathbb{R}^d} |k^{\mathbf{w}}(\mathbf{u}')| \|\mathbf{u}'\|^2 d\mathbf{u}' \end{aligned}$$

That this final bound is finite follows from the requirement that  $\int_{\mathbb{R}} (|\overline{\widehat{\psi}(\xi)} \widehat{\phi}(\xi)| / |\xi|^{d+2}) d\xi < \infty$  in Definition 2, together with the assumption that  $\widehat{p}_{\mathbf{w}}$  has finite second moment, which ensures the finiteness of  $\int_{\mathbb{R}^d} |k^{\mathbf{w}}(\mathbf{u}')| \|\mathbf{u}'\|^2 d\mathbf{u}'$ . Next we verify (26). From (23) and the change of variables  $\mathbf{u}' = \xi \mathbf{u}$ ,

$$\begin{aligned} \int_{\mathbb{R}^d} \tilde{k}(\mathbf{v}, \mathbf{v} - \mathbf{u}) d\mathbf{u} &= (2\pi)^{\frac{d}{2}} \int_{\mathbb{R}^d} \int_{\mathbb{R}} \overline{\widehat{\psi}(\xi)} \widehat{\phi}(\xi_1) k^{\mathbf{w}}(\xi \mathbf{u}) d\xi d\mathbf{u} \\ &= (2\pi)^{\frac{d}{2}} \int_{\mathbb{R}} \frac{\overline{\widehat{\psi}(\xi)} \widehat{\phi}(\xi)}{|\xi|^d} d\xi \int_{\mathbb{R}^d} k^{\mathbf{w}}(\mathbf{u}') d\mathbf{u}' = 1, \end{aligned}$$

where the final inequality used the facts that  $\int_{\mathbb{R}^d} k^{\mathbf{w}}(\mathbf{u}') d\mathbf{u}' = \int_{\mathbb{R}^d} |k^{\mathbf{w}}(\mathbf{u}')| d\mathbf{u}' = 1$  by construction and  $(2\pi)^{\frac{d}{2}} \int_{\mathbb{R}} \frac{\widehat{\psi}(\xi)\widehat{\phi}(\xi)}{|\xi|^d} d\xi = 1$  from Definition 2. Finally we verify (25). From (23) and the change of variables  $\mathbf{u}' = \xi\mathbf{u}$ ,

$$\begin{aligned} \int_{\mathbb{R}^d} \widetilde{k}(\mathbf{v}, \mathbf{v} - \mathbf{u}) \mathbf{u}^\alpha d\mathbf{u} &= (2\pi)^{\frac{d}{2}} \int_{\mathbb{R}^d} \int_{\mathbb{R}} \overline{\widehat{\psi}(\xi_1)\widehat{\phi}(\xi_1)} k^{\mathbf{w}}(\xi\mathbf{u}) d\xi \mathbf{u}^\alpha d\mathbf{u} \\ &= \underbrace{(2\pi)^{\frac{d}{2}} \int_{\mathbb{R}} \frac{\widehat{\psi}(\xi)\widehat{\phi}(\xi)}{|\xi|^d} \frac{1}{\xi^{|\alpha|}} d\xi}_{(1)} \underbrace{\int_{\mathbb{R}^d} k^{\mathbf{w}}(\mathbf{u}') \mathbf{u}'^\alpha d\mathbf{u}'}_{(2)}. \end{aligned}$$

The term (1) is finite as a consequence of the assumption  $\int_{\mathbb{R}} \frac{|\widehat{\psi}(\xi)\widehat{\phi}(\xi)|}{|\xi|^{d+2}} d\xi < \infty$  in Definition 2. For the term (2), note that  $\mathbf{u} \mapsto \mathbf{u}^\alpha$  is odd function whenever  $|\alpha| = 1$ . On the other hand,  $k^{\mathbf{w}}$  is even function since  $k^{\mathbf{w}}$  is the Fourier transform of an even probability density  $p_{\mathbf{w}}$ . Thus the function  $\mathbf{u} \mapsto k^{\mathbf{w}}(\mathbf{u}) \mathbf{u}^\alpha$ , which is given by the product of even and odd functions, is odd. For any integrable odd function  $h : \mathbb{R}^d \rightarrow \mathbb{R}$ ,  $\int_{\mathbb{R}^d} h(\mathbf{u}) d\mathbf{u} = 0$  holds as  $\int_{\mathbb{R}_+^d} h(\mathbf{u}) d\mathbf{u} = -\int_{\mathbb{R}_-^d} h(\mathbf{u}) d\mathbf{u}$  where  $\mathbb{R}_+^d$  and  $\mathbb{R}_-^d$  are positive and negative half Euclidean space. The function  $\mathbf{u} \mapsto k^{\mathbf{w}}(\mathbf{u}) \mathbf{u}^\alpha$  is integrable since  $k^{\mathbf{w}}$  has the finite second moment and  $|\alpha| = 1$ . This implies  $\int_{\mathbb{R}^d} \widetilde{k}(\mathbf{v}, \mathbf{v} - \mathbf{u}) \mathbf{u}^\alpha d\mathbf{u} = \mathbf{0}$ .

Thus  $\widetilde{k}$  satisfies the condition in Proposition 1 and, with  $K_h[f] = R_\sigma^* R^{(w)} f$ , we obtain for some  $C_1 > 0$  depending on  $\widetilde{k}$  but not  $f$ ,

$$\sup_{\mathbf{x} \in \mathbb{R}^d} |f(\mathbf{x}) - R_\sigma^* R^{(w)} f(\mathbf{x})| \leq C_1 M_1(f) \epsilon_{\mathbf{w}}. \quad (28)$$

**Bounding (\*\*):** A bound on (\*\*) makes use of Auxiliary Lemma 1. Define  $k_\epsilon : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  by

$$k_\epsilon(\mathbf{x}, \mathbf{x}') := (2\pi)^{\frac{1}{2}} \|\widehat{p}_{\mathbf{w}}\|_{L^1}^{-1} \|\widehat{p}_b\|_{L^1}^{-1} \int_{\mathbb{R}} \int_{\mathbb{R}^d} \psi(\mathbf{w} \cdot \mathbf{x}' + b) \phi(\mathbf{w} \cdot \mathbf{x} + b) p_{\mathbf{w}}^\epsilon(\mathbf{w}) p_b^\epsilon(b) d\mathbf{w} db. \quad (29)$$

so that  $R_\sigma^* R f(\mathbf{x}) = \int_{\mathbb{R}^d} f(\mathbf{x}') k_\epsilon(\mathbf{x}, \mathbf{x}') d\mathbf{x}'$  and recall that  $R_\sigma^* R^{(w)} f(\mathbf{x}) = \int_{\mathbb{R}^d} f(\mathbf{x}') \widetilde{k}_\epsilon(\mathbf{x}, \mathbf{x}') d\mathbf{x}'$  where  $\widetilde{k}_\epsilon$  is defined by (24). The second error term is

$$\begin{aligned} \sup_{\mathbf{x} \in \mathbb{R}^d} |R_\sigma^* R^{(w)} f(\mathbf{x}) - R_\sigma^* R f(\mathbf{x})| &= \sup_{\mathbf{x} \in \mathbb{R}^d} \left| \int_{\mathbb{R}^d} f(\mathbf{x}') \widetilde{k}_\epsilon(\mathbf{x}, \mathbf{x}') d\mathbf{x}' - \int_{\mathbb{R}^d} f(\mathbf{x}') k_\epsilon(\mathbf{x}, \mathbf{x}') d\mathbf{x}' \right| \\ &= \sup_{\mathbf{x} \in \mathbb{R}^d} \left| \int_{\mathbb{R}^d} f(\mathbf{x}') (\widetilde{k}_\epsilon(\mathbf{x}, \mathbf{x}') - k_\epsilon(\mathbf{x}, \mathbf{x}')) d\mathbf{x}' \right|. \end{aligned}$$

Let  $\Delta k(\mathbf{x}, \mathbf{x}') := \widetilde{k}_\epsilon(\mathbf{x}, \mathbf{x}') - k_\epsilon(\mathbf{x}, \mathbf{x}')$  for short-hand. By the definition of  $\widetilde{k}_\epsilon$  and  $k_\epsilon$ ,

$$\Delta k(\mathbf{x}, \mathbf{x}') = \|\widehat{p}_{\mathbf{w}}\|_{L^1}^{-1} \int_{\mathbb{R}} \int_{\mathbb{R}^d} \psi(\mathbf{w} \cdot \mathbf{x}' + b) \phi(\mathbf{w} \cdot \mathbf{x} + b) p_{\mathbf{w}}^\epsilon(\mathbf{w}) \left\{ 1 - (2\pi)^{\frac{1}{2}} \|\widehat{p}_b\|_{L^1}^{-1} p_b^\epsilon(b) \right\} d\mathbf{w} db.$$

Next we upper bound  $\Delta k(\mathbf{x}, \mathbf{x}')$ . Substituting the identity established in Lemma 1 yields

$$\begin{aligned} \Delta k(\mathbf{x}, \mathbf{x}') &= \epsilon_b(2\pi)^{\frac{1}{2}} \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \|\widehat{p_b}\|_{L^1}^{-1} \int_{\mathbb{R}} \int_{\mathbb{R}^d} \psi(\mathbf{w} \cdot \mathbf{x}' + b) \phi(\mathbf{w} \cdot \mathbf{x} + b) p_{\mathbf{w}}^\epsilon(\mathbf{w}) \int_0^1 b \partial p_b(t\epsilon_b b) dt d\mathbf{w} db \\ &= \epsilon_b(2\pi)^{\frac{1}{2}} \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \|\widehat{p_b}\|_{L^1}^{-1} \int_0^1 \int_{\mathbb{R}^d} \underbrace{\int_{\mathbb{R}} b \phi(\mathbf{w} \cdot \mathbf{x} + b) \psi(\mathbf{w} \cdot \mathbf{x}' + b) \partial p_b(t\epsilon_b b) db}_{(3)} p_{\mathbf{w}}^\epsilon(\mathbf{w}) d\mathbf{w} dt. \end{aligned}$$

By Assumption 4,  $p_b$  has bounded first derivative. Thus:

$$|(3)| \leq \|\partial p_b\|_{L^\infty} \int_{\mathbb{R}} |b \phi(\mathbf{w} \cdot \mathbf{x} + b) \psi(\mathbf{w} \cdot \mathbf{x}' + b)| db'$$

By a change of variable  $b' = \mathbf{w} \cdot \mathbf{x}' + b$ ,

$$\begin{aligned} |(3)| &\leq \|\partial p_b\|_{L^\infty} \int_{\mathbb{R}} |b' - \mathbf{w} \cdot \mathbf{x}'| |\phi(\mathbf{w} \cdot (\mathbf{x} - \mathbf{x}') + b') \psi(b')| db' \\ &\leq \|\partial p_b\|_{L^\infty} \int_{\mathbb{R}} (|b'| + \|\mathbf{w}\|_2 \|\mathbf{x}'\|_2) |\phi(\mathbf{w} \cdot (\mathbf{x} - \mathbf{x}') + b') \psi(b')| db' \end{aligned}$$

where the triangle inequality and Cauchy-Schwartz inequality have been applied. Then it is easy to see

$$|(3)| \leq \|\partial p_b\|_{L^\infty} (1 + \|\mathbf{x}'\|_2) (1 + \|\mathbf{w}\|_2) \int_{\mathbb{R}} (1 + |b'|) |\phi(\mathbf{w} \cdot (\mathbf{x} - \mathbf{x}') + b') \psi(b')| db' \quad (30)$$

By the assumption  $\phi$  is bounded,

$$|(3)| \leq \|\partial p_b\|_{L^\infty} \|\phi\|_{L^\infty} (1 + \|\mathbf{x}'\|_2) (1 + \|\mathbf{w}\|_2) \int_{\mathbb{R}} (1 + |b'|) |\psi(b')| db' \quad (31)$$

Since  $\psi$  is Schwartz function, the integral is finite. Let  $C_\psi := \int_{\mathbb{R}} (1 + |b|) |\psi(b)| db$  to see  $|(3)| \leq \|\partial p_b\|_{L^\infty} \|\phi\|_{L^\infty} C_\psi C_{p_{\mathbf{w}}} (1 + \|\mathbf{x}'\|_2) (1 + \|\mathbf{w}\|_2)$ . From this upper bound of  $|(3)|$  we have

$$\begin{aligned} \Delta k(\mathbf{x}, \mathbf{x}') &= \epsilon_b(2\pi)^{\frac{1}{2}} \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \|\widehat{p_b}\|_{L^1}^{-1} \int_0^1 \int_{\mathbb{R}^d} |(3)| p_{\mathbf{w}}^\epsilon(\mathbf{w}) d\mathbf{w} dt \\ &\leq \epsilon_b(2\pi)^{\frac{1}{2}} \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \|\widehat{p_b}\|_{L^1}^{-1} \|\partial p_b\|_{L^\infty} \|\phi\|_{L^\infty} C_\psi C_{p_{\mathbf{w}}} (1 + \|\mathbf{x}'\|_2) \int_{\mathbb{R}^d} (1 + \|\mathbf{w}\|_2) p_{\mathbf{w}}^\epsilon(\mathbf{w}) d\mathbf{w} \int_0^1 dt \end{aligned}$$

For the integral  $\int_{\mathbb{R}^d} (1 + \|\mathbf{w}\|_2) p_{\mathbf{w}}^\epsilon(\mathbf{w}) d\mathbf{w}$ , by a change of variable  $\mathbf{w}' = \epsilon_{\mathbf{w}} \mathbf{w}$ ,

$$\int_{\mathbb{R}^d} (1 + \|\mathbf{w}\|_2) p_{\mathbf{w}}^\epsilon(\mathbf{w}) d\mathbf{w} = \int_{\mathbb{R}^d} \left( \frac{1}{\epsilon_{\mathbf{w}}^d} + \frac{\|\mathbf{w}'\|_2}{\epsilon_{\mathbf{w}}^{d+1}} \right) p_{\mathbf{w}}(\mathbf{w}') d\mathbf{w}'.$$

Recall that  $p_{\mathbf{w}}$  was assumed to have finite second moment. Let  $C_{p_{\mathbf{w}}} := \max(1, \int_{\mathbb{R}^d} \|\mathbf{w}\|_2 p_{\mathbf{w}}(\mathbf{w}) d\mathbf{w})$  to see

$$\int_{\mathbb{R}^d} \left( \frac{1}{\epsilon_{\mathbf{w}}^d} + \frac{\|\mathbf{w}'\|_2}{\epsilon_{\mathbf{w}}^{d+1}} \right) p_{\mathbf{w}}(\mathbf{w}') d\mathbf{w}' \leq C_{p_{\mathbf{w}}} \left( \frac{1}{\epsilon_{\mathbf{w}}^d} + \frac{1}{\epsilon_{\mathbf{w}}^{d+1}} \right) = C_{p_{\mathbf{w}}} \frac{1}{\epsilon_{\mathbf{w}}^d} \left( 1 + \frac{1}{\epsilon_{\mathbf{w}}} \right).$$



Plugging this upper bound in  $\Delta k(\mathbf{x}, \mathbf{x}')$ ,

$$\Delta k(\mathbf{x}, \mathbf{x}') \leq (2\pi)^{\frac{1}{2}} \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \|\widehat{p_b}\|_{L^1}^{-1} \|\partial p_b\|_{L^\infty} \|\phi\|_{L^\infty} C_\psi C_{p_{\mathbf{w}}} (1 + \|\mathbf{x}'\|_2) \frac{1}{\epsilon_{\mathbf{w}}^d} \left(1 + \frac{1}{\epsilon_{\mathbf{w}}}\right) \epsilon_b.$$

The original error term is then bounded as

$$\begin{aligned} (**) &= \sup_{\mathbf{x} \in \mathbb{R}^d} \left| \int_{\mathbb{R}^d} f(\mathbf{x}') \Delta k(\mathbf{x}, \mathbf{x}') d\mathbf{x}' \right| \\ &\leq (2\pi)^{\frac{1}{2}} \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \|\widehat{p_b}\|_{L^1}^{-1} \|\partial p_b\|_{L^\infty} \|\phi\|_{L^\infty} C_\psi C_{p_{\mathbf{w}}} \int_{\mathbb{R}^d} (1 + \|\mathbf{x}'\|_2) |f(\mathbf{x}')| d\mathbf{x}' \frac{1}{\epsilon_{\mathbf{w}}^d} \left(1 + \frac{1}{\epsilon_{\mathbf{w}}}\right) \epsilon_b. \end{aligned}$$

From the assumption of  $f$  in Theorem 2,  $B_1(f) = \int_{\mathbb{R}^d} (1 + \|\mathbf{x}\|_2) |f(\mathbf{x})| d\mathbf{x} < \infty$ . Setting  $C_2 := (2\pi)^{\frac{1}{2}} \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \|\widehat{p_b}\|_{L^1}^{-1} \|\partial p_b\|_{L^\infty} \|\phi\|_{L^\infty} C_\psi C_{p_{\mathbf{w}}}$  gives

$$(**) \leq C_2 B_1(f) \frac{1}{\epsilon_{\mathbf{w}}^d} \left(1 + \frac{1}{\epsilon_{\mathbf{w}}}\right) \epsilon_b. \quad (32)$$

**Overall Bound:** Substituting the results of (28) and (32) into (18), for some  $C > 0$ ,

$$\sup_{\mathbf{x} \in \mathbb{R}^d} |f(\mathbf{x}) - R_\sigma^* R f(\mathbf{x})| \leq C \max(M_1(f), B_1(f)) \left\{ \epsilon_{\mathbf{w}} + \frac{1}{\epsilon_{\mathbf{w}}^d} \left(1 + \frac{1}{\epsilon_{\mathbf{w}}}\right) \epsilon_b \right\} \quad (33)$$

where  $C$  only depends on  $\phi, \psi, p_{\mathbf{w}}, p_b$ . Setting  $\sigma_{\mathbf{w}} = \epsilon_{\mathbf{w}}^{-1}$  and  $\sigma_b = \epsilon_b^{-1}$ , the main convergence result is obtained.  $\blacksquare$

### A.1.3 AUXILIARY LEMMA 1

The following technical lemma was exploited in the proof of Theorem 2:

**Lemma 1.** *For  $p_b$  in the setting of Assumption 4, we have that*

$$1 - (2\pi)^{\frac{1}{2}} \|\widehat{p_b}\|_{L^1}^{-1} p_b^\epsilon(b) = \epsilon_b (2\pi)^{\frac{1}{2}} \|\widehat{p_b}\|_{L^1}^{-1} \int_0^1 b \partial p_b(t \epsilon_b b) dt.$$

**Proof** The result will be established by proving (a)  $1 = (2\pi)^{\frac{1}{2}} \|\widehat{p_b}\|_{L^1}^{-1} p_b^\epsilon(0)$  and (b)  $p_b(0) - p_b(\epsilon_b b) = \epsilon_b \int_0^1 b \partial p_b(t \epsilon_b b) dt$ , which are algebraically seen to imply the stated result.

**Part (a):** Recall that the *Fourier inversion*  $g(\mathbf{x}) = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}} \widehat{g}(\boldsymbol{\xi}) e^{i\boldsymbol{\xi} \cdot \mathbf{x}} d\boldsymbol{\xi}$  holds for any function  $g \in L^1(\mathbb{R}^d)$  s.t.  $\widehat{g} \in L^1(\mathbb{R}^d)$ . We use the fact  $g(\mathbf{0}) = (2\pi)^{-\frac{d}{2}} \|\widehat{g}\|_{L^1}$  for  $g \in L^1(\mathbb{R}^d)$  s.t.  $\widehat{g} \in L^1(\mathbb{R}^d)$  and  $\widehat{g}$  is positive, which is obtained by substituting  $\mathbf{x} = \mathbf{0}$  into the Fourier inversion  $(2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \widehat{g}(\mathbf{w}) e^{i\mathbf{w} \cdot \mathbf{0}} d\mathbf{w} = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} \widehat{g}(\mathbf{w}) d\mathbf{w}$ . Recall that  $p_b^\epsilon(b) = p_b(\epsilon_b b)$ . From standard properties of the Fourier transform (Grafakos, 2000, p.109,113), the Fourier transform of  $b \mapsto p_b(\epsilon_b b)$  is given as  $\xi \mapsto \frac{1}{\epsilon_b} \widehat{p_b} \left( \frac{\xi}{\epsilon_b} \right)$  and  $\widehat{p_b}$  is positive by the assumption. Hence,  $p_b^\epsilon(0) = (2\pi)^{-\frac{1}{2}} \|\widehat{p_b^\epsilon}\|_{L^1} = (2\pi)^{-\frac{1}{2}} \left\| \frac{1}{\epsilon_b} \widehat{p_b} \left( \frac{\cdot}{\epsilon_b} \right) \right\|_{L^1}$ . Since the  $L^1(\mathbb{R})$  norm is invariant

to the scaling of function i.e.  $\left\| \frac{1}{\epsilon_b} \widehat{p}_b \left( \frac{\cdot}{\epsilon_b} \right) \right\|_{L^1} = \|\widehat{p}_b\|_{L^1}$ , we obtain  $p_b^\epsilon(0) = (2\pi)^{-\frac{1}{2}} \|\widehat{p}_b\|_{L^1}$  as required.

**Part (b):** We use the fact that the equation  $g(y) - g(x) = \int_0^1 (y-x) \partial g(x-t(y-x)) dt$  holds for  $g \in C^1(\mathbb{R})$  (Gine and Nickl, 2015, p.302,304) to see that

$$p_b(0) - p_b(\epsilon_b b) = \int_0^1 (-\epsilon_b b) \partial p_b((1-t')\epsilon_b b) dt' = \epsilon_b \int_0^1 b \partial p_b(t\epsilon_b b) dt$$

where change of variable  $t = 1 - t'$  is applied. This holds since  $p_b \in C^1(\mathbb{R})$ . ■

## A.2 Proof of Theorem 1

This section is dedicated to the proof of Theorem 1. It is divided into three parts; in Appendix A.2.1 we state technical lemmas that will be useful; in Appendix A.2.2 we state and prove an intermediate result concerning the discretised ridgelet transform, then in Appendix A.2.3 we present the proof of Theorem 1.

### A.2.1 TECHNICAL LEMMAS

The following technical lemmas will be useful for the proof of Theorem 1.

**Lemma 2.** *Let  $\mathcal{X}$  be a bounded subset of  $\mathbb{R}^d$ . Let  $g : \mathcal{X} \times \mathbb{R}^p \rightarrow \mathbb{R}$  be such that  $g(\mathbf{x}, \cdot) : \mathbb{R}^p \rightarrow \mathbb{R}$  are measurable for all  $\mathbf{x} \in \mathcal{X}$ . Let  $\boldsymbol{\theta}, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$  be independent samples from a distribution  $\mathbb{P}$  on  $\mathbb{R}^p$ . Assume that there exists a measurable function  $G : \mathbb{R}^p \rightarrow \mathbb{R}$  such that  $\mathbb{E}[G(\boldsymbol{\theta})^2] < \infty$  and*

$$|g(\mathbf{x}, \boldsymbol{\theta}) - g(\mathbf{x}', \boldsymbol{\theta})| \leq G(\boldsymbol{\theta}) \|\mathbf{x} - \mathbf{x}'\|_2 \text{ for all } \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \quad (34)$$

Then

$$\mathbb{E} \left[ \sup_{\mathbf{x} \in \mathcal{X}} \left| \frac{1}{n} \sum_{i=1}^n g(\mathbf{x}, \boldsymbol{\theta}_i) - \mathbb{E}[g(\mathbf{x}, \boldsymbol{\theta})] \right| \right] \leq \frac{C}{\sqrt{n}} \sqrt{\mathbb{E}[G(\boldsymbol{\theta})^2]}.$$

where  $C$  is a constant that depends only on  $\mathcal{X}$ .

**Proof** Let  $g_{\mathbf{x}} := g(\mathbf{x}, \cdot)$  for shorthand and let  $\mathcal{G} := \{g_{\mathbf{x}} | \mathbf{x} \in \mathcal{X}\}$ . For any  $g_{\mathbf{x}}, g_{\mathbf{x}'} \in \mathcal{G}$ , define a (random) pseudo metric  $\rho_n(g_{\mathbf{x}}, g_{\mathbf{x}'} ) := \sqrt{\frac{1}{n} \sum_{i=1}^n (g_{\mathbf{x}}(\boldsymbol{\theta}_i) - g_{\mathbf{x}'}(\boldsymbol{\theta}_i))^2}$  and the diameter  $D_n := \sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} \rho_n(g_{\mathbf{x}}, g_{\mathbf{x}'})$ . Let  $N(\mathcal{G}, \rho_n, \epsilon)$  denotes the *covering number* of the set  $\mathcal{G}$  by  $\epsilon$ -ball under the metric  $\rho_n$  (Gine and Nickl, 2015, p.41). Then by (Gine and Nickl, 2015, Theorem 3.5.1, Remark 3.5.2, p.185),

$$\mathbb{E} \left[ \sup_{\mathbf{x} \in \mathcal{X}} \left| \frac{1}{n} \sum_{i=1}^n g_{\mathbf{x}}(\boldsymbol{\theta}_i) - \mathbb{E}[g_{\mathbf{x}}(\cdot)] \right| \right] \leq \frac{8\sqrt{2}}{\sqrt{n}} \underbrace{\mathbb{E} \left[ \int_0^{D_n} \sqrt{\log 2N(\mathcal{G}, \rho_n, \epsilon)} d\epsilon \right]}_{(*)}$$

Here we reduced the assumption  $0 \in \mathcal{G}$  by the discussion in (Wainwright, 2019, p.135). Let  $\|G\|_{\rho_n} = \sqrt{\frac{1}{n} \sum_{i=1}^n G(\boldsymbol{\theta}_i)^2}$ . By (Kosorok, 2008, Lemma 9.18, p.166) and (van der Vaart,

1998, Example 19.7, p.271), the covering number is bounded as  $N(\mathcal{G}, \rho_n, \epsilon) \leq \left(\frac{K\|G\|_{\rho_n}}{\epsilon}\right)^d$  where  $K$  is a constant depending only on  $\mathcal{X}$ . Hence  $(*) \leq \sqrt{d} \int_0^{D_n} \sqrt{\log \frac{2K\|G\|_{\rho_n}}{\epsilon}} d\epsilon$ . By Cauchy-Schwartz inequality,

$$\sqrt{d} \int_0^{D_n} \sqrt{\log \frac{2K\|G\|_{\rho_n}}{\epsilon}} d\epsilon \leq \sqrt{dD_n} \sqrt{\int_0^{D_n} \log \frac{2K\|G\|_{\rho_n}}{\epsilon} d\epsilon}.$$

By calculating the integral,

$$(*) \leq \sqrt{dD_n} \sqrt{D_n \left(1 + \log \frac{2K\|G\|_{\rho_n}}{D_n}\right)} = \sqrt{dD_n} \sqrt{1 + \log \frac{2K\|G\|_{\rho_n}}{D_n}}.$$

By the preceding assumption (34),

$$D_n = \sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} \sqrt{\frac{1}{n} \sum_{i=1}^n (g_{\mathbf{x}}(\boldsymbol{\theta}_i) - g_{\mathbf{x}'}(\boldsymbol{\theta}_i))^2} \leq \sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} \sqrt{\frac{1}{n} \sum_{i=1}^n G(\boldsymbol{\theta}_i)^2 \|\mathbf{x} - \mathbf{x}'\|_2^2} = R\|G\|_{\rho_n}.$$

where  $R = \sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} \|\mathbf{x} - \mathbf{x}'\|_2$ . We can set  $K$  so that  $R \leq 2K$  without loss of generality, then we have  $\frac{2K\|G\|_{\rho_n}}{D_n} \geq 1 \Leftrightarrow \log \frac{2K\|G\|_{\rho_n}}{D_n} \geq 0$  which implies

$$\sqrt{dD_n} \sqrt{1 + \log \frac{2K\|G\|_{\rho_n}}{D_n}} \leq \sqrt{dD_n} \left(1 + \log \frac{2K\|G\|_{\rho_n}}{D_n}\right).$$

By the inequality  $1 + \log z \leq z$  for all  $z > 0$ , we have  $1 + \log \frac{2K\|G\|_{\rho_n}}{D_n} \leq \frac{2K\|G\|_{\rho_n}}{D_n}$  and

$$(*) \leq 2\sqrt{d}K\|G\|_{\rho_n}.$$

Then by Jensen' inequality,

$$\mathbb{E}[(*)] = 2\sqrt{d}K\mathbb{E} \left[ \sqrt{\frac{1}{n} \sum_{i=1}^n G(\boldsymbol{\theta}_i)^2} \right] \leq 2\sqrt{d}K \sqrt{\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n G(\boldsymbol{\theta}_i)^2 \right]} = 2\sqrt{d}K \sqrt{\mathbb{E}[G(\boldsymbol{\theta})^2]}.$$

This completes the proof. ■

**Lemma 3.** *For any  $b$ -uniformly bounded class of function  $\mathcal{F}$  and any integer  $n \geq 1$ , we have, with probability at least  $1 - \delta$ ,*

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \leq \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \right] + \frac{b\sqrt{2 \log \delta^{-1}}}{\sqrt{n}}.$$

**Proof** From equation (4.16) in Wainwright (2019), with probability at least  $1 - \exp\left(-\frac{n\delta'^2}{2b^2}\right)$

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \leq \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \right] + \delta'.$$

Setting  $\delta' = \frac{b\sqrt{2\log\delta^{-1}}}{\sqrt{n}}$  yields the result.  $\blacksquare$

**Lemma 4.** *Let  $S$  be a positive constant and  $h \in C^1([-S, S]^d)$ . For grid points  $(\mathbf{x}_i)_{i=1}^D$  on  $[-S, S]^d$  corresponding to a Cartesian product of left endpoint rules, we have that*

$$\left| \int_{[-S, S]^d} h(\mathbf{x}) d\mathbf{x} - \frac{(2S)^d}{D} \sum_{i=1}^D h(\mathbf{x}_i) \right| \leq \frac{C}{D^{\frac{1}{d}}} \max_{|\alpha|=1} \sup_{\mathbf{x} \in [-S, S]^d} |\partial^\alpha h(\mathbf{x})|. \quad (35)$$

where  $C$  is a constant independent of  $h$ .

**Proof** Let  $r := 2S/\sqrt[d]{D}$ . For  $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,d})$ , let  $R_i := [x_{i,1}, x_{i,1}+r] \times \dots \times [x_{i,d}, x_{i,d}+r]$  for  $i = 1, \dots, D$ . Since  $(\mathbf{x}_i)_{i=1}^D$  are grid points on  $[-S, S]^d$  corresponding to a Cartesian product of left endpoint rules, the domain  $[-S, S]^d$  can be decomposed as  $[-S, S]^d = R_1 \oplus \dots \oplus R_D$ , meaning

$$\int_{[-S, S]^d} h(\mathbf{x}) d\mathbf{x} = \sum_{i=1}^D \int_{R_i} h(\mathbf{x}) d\mathbf{x}.$$

Denote the original error in (35) by  $(*)$ . Then noting  $r^d = \int_{R_i} d\mathbf{x}$ ,

$$(*) = \left| \sum_{i=1}^D \int_{R_i} h(\mathbf{x}) d\mathbf{x} - r^d \sum_{i=1}^D h(\mathbf{x}_i) \right| = \left| \sum_{i=1}^D \int_{R_i} (h(\mathbf{x}) - h(\mathbf{x}_i)) d\mathbf{x} \right|$$

By the mean value theorem, there exists  $\mathbf{x}_i^*$  for  $i = 1, \dots, D$  such that

$$(*) = \left| \sum_{i=1}^D \int_{R_i} \nabla h(\mathbf{x}_i^*) \cdot (\mathbf{x} - \mathbf{x}_i) d\mathbf{x} \right|$$

where  $\nabla h$  is the gradient vector of  $h$ . Calculating the integral and taking supremum of  $\nabla h$

$$(*) \leq \sum_{i=1}^D \frac{d}{2} r^{d+1} \max_{|\alpha|=1} \sup_{\mathbf{x} \in [-S, S]^d} |\partial^\alpha h(\mathbf{x})| = \frac{d}{2} D r^{d+1} \max_{|\alpha|=1} \sup_{\mathbf{x} \in [-S, S]^d} |\partial^\alpha h(\mathbf{x})|.$$

Substituting  $r = 2S/\sqrt[d]{D}$  and setting  $C := d(2S)^{d+1}/2$  concludes the proof.  $\blacksquare$

**Lemma 5.** *For a non-negative random variable  $X$  such that  $\mathbb{E}[X] < \infty$ , it holds with probability at least  $1 - \delta$  that  $X \leq \mathbb{E}[X]/\delta$ .*

**Proof** From the *Markov inequality* we have that  $\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$ . Taking a complement of the probability and setting  $\delta = \frac{\mathbb{E}[X]}{t}$ , we obtain the result.  $\blacksquare$

## A.2.2 AN INTERMEDIATE RIDGELET RECONSTRUCTION RESULT

The aim of this section is to establish an analogue of Theorem 2 that holds when the ridgelet operator  $R_\sigma^* R$  is discretised using cubature rules, as in (12). The purpose of Theorem 3 is to guarantee an accurate reconstruction with high probability. This will be central to the proof of Theorem 1. We introduce the generalisation of Assumption 3.

**Assumption 5.** For densities  $p_{\mathbf{w}}$  on  $\mathbb{R}^d$  and  $p_b$  on  $\mathbb{R}$  satisfying Assumption 4, define scaled densities  $p_{\mathbf{w},\sigma}(\mathbf{w}) := \sigma_{\mathbf{w}}^{-d} p_{\mathbf{w}}(\sigma_{\mathbf{w}}^{-1} \mathbf{w})$  and  $p_{b,\sigma}(b) := \sigma_b^{-1} p_b(\sigma_b^{-1} b)$  with scaling constants  $0 < \sigma_{\mathbf{w}} < \infty$  and  $0 < \sigma_b < \infty$ . The cubature nodes  $\{\mathbf{w}_i, b_i\}_{i=1}^N$  are independently sampled from  $p_{\mathbf{w},\sigma} \times p_{b,\sigma}$  and  $v_i := Z/N$  for all  $i = 1, \dots, N$ , where  $Z := (2\pi)^{\frac{1}{2}} \sigma_{\mathbf{w}}^d \sigma_b \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \|\widehat{p_b}\|_{L^1}^{-1}$ .

Note that Assumption 3 is recovered by setting Gaussians with variances  $\sigma_{\mathbf{w}}^2 \mathbf{I}_{d \times d}$  and  $\sigma_b^2$  to the scaled densities  $p_{\mathbf{w},\sigma}$  and  $p_{b,\sigma}$ . Recall  $M_1(f) = \max_{|\alpha| \leq 1} \sup_{\mathbf{x} \in \mathbb{R}^d} |\partial^\alpha f(\mathbf{x})|$  and  $M_1^*(f) = \max_{|\alpha| \leq 1} \sup_{\mathbf{x} \in [-S, S]^d} |\partial^\alpha f(\mathbf{x})|$ .

**Theorem 3.** Let  $I_{\sigma,D,N}$  is given by (12) under Assumption 2 and 5. Further, assume  $\phi$  is  $L_\phi$ -Lipschitz continuous. For any  $f \in C^1(\mathbb{R}^d)$  with  $M_1^*(f) < \infty$ , with probability at least  $1 - \delta$ ,

$$\begin{aligned} & \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x}) - I_{\sigma,D,N} f(\mathbf{x})| \\ & \leq C M_1^*(f) \left\{ \frac{1}{\sigma_{\mathbf{w}}} + \frac{\sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)}{\sigma_b} + \frac{\sigma_b \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)}{D^{\frac{1}{d}}} + \frac{\sigma_b \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + \sqrt{\log \delta^{-1}})}{\sqrt{N}} \right\} \end{aligned}$$

where  $C$  is a constant that may depend on  $\mathcal{X}, \mathbb{1}, \phi, \psi, p_{\mathbf{w}}, p_b$  but does not depend on  $f, \sigma_{\mathbf{w}}, \sigma_b, \delta$ .

**Proof** Under Assumption 2, specifically  $u_j = (2S)^d D^{-1} \mathbb{1}(\mathbf{x}_j)$  and  $v_i = Z/N$ , we have that

$$I_{\sigma,D,N} f(\mathbf{x}) = \sum_{i=1}^N \frac{Z}{N} \left( \frac{(2S)^d}{D} \sum_{j=1}^D \mathbb{1}(\mathbf{x}_j) f(\mathbf{x}_j) \psi(\mathbf{w}_i \cdot \mathbf{x}_j + b_i) \right) \phi(\mathbf{w}_i \cdot \mathbf{x} + b_i).$$

and we formally define

$$\begin{aligned} I_\sigma f(\mathbf{x}) & := \int_{\mathbb{R}} \int_{\mathbb{R}^d} \left( \int_{\mathbb{R}^d} \mathbb{1}(\mathbf{x}') f(\mathbf{x}') \psi(\mathbf{w} \cdot \mathbf{x}' + b) d\mathbf{x}' \right) \phi(\mathbf{w} \cdot \mathbf{x} + b) Z p_{\mathbf{w},\sigma}(\mathbf{w}) p_{b,\sigma}(b) d\mathbf{w} db, \\ I_{\sigma,D} f(\mathbf{x}) & := \int_{\mathbb{R}} \int_{\mathbb{R}^d} \left( \frac{(2S)^d}{D} \sum_{j=1}^D \mathbb{1}(\mathbf{x}'_j) f(\mathbf{x}'_j) \psi(\mathbf{w} \cdot \mathbf{x}'_j + b) \right) \phi(\mathbf{w} \cdot \mathbf{x} + b) Z p_{\mathbf{w},\sigma}(\mathbf{w}) p_{b,\sigma}(b) d\mathbf{w} db. \end{aligned}$$

The error will be decomposed by the triangle inequality,

$$\begin{aligned} & \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x}) - I_{\sigma,D,N} f(\mathbf{x})| \\ & \leq \underbrace{\sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x}) - I_\sigma f(\mathbf{x})|}_{(*)} + \underbrace{\sup_{\mathbf{x} \in \mathcal{X}} |I_\sigma f(\mathbf{x}) - I_{\sigma,D} f(\mathbf{x})|}_{(**)} + \underbrace{\sup_{\mathbf{x} \in \mathcal{X}} |I_{\sigma,D} f(\mathbf{x}) - I_{\sigma,D,N} f(\mathbf{x})|}_{(***)}. \quad (36) \end{aligned}$$

**Bounding (\*)**: Let  $f_{\text{clip}}(\mathbf{x}) := \mathbb{1}(\mathbf{x})f(\mathbf{x})$ . Since  $f_{\text{clip}}(\mathbf{x}) = f(\mathbf{x})$  for all  $\mathbf{x} \in \mathcal{X}$  and our construction ensures that  $I_\sigma f(\mathbf{x}) = R_\sigma^* R f_{\text{clip}}(\mathbf{x})$ , we have that

$$(*) = \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x}) - I_\sigma f(\mathbf{x})| = \sup_{\mathbf{x} \in \mathcal{X}} |f_{\text{clip}}(\mathbf{x}) - R_\sigma^* R f_{\text{clip}}(\mathbf{x})|.$$

In order to apply Theorem 2 for (\*), the two quantities  $M_1(f_{\text{clip}})$  and  $B(f_{\text{clip}})$  must be shown to be finite. For the first quantity, we have

$$\begin{aligned} M_1(f_{\text{clip}}) &\leq \sup_{\mathbf{x} \in \mathbb{R}^d} |(\mathbb{1}f)(\mathbf{x})| + \max_{|\alpha|=1} \sup_{\mathbf{x} \in \mathbb{R}^d} |\partial^\alpha (\mathbb{1}f)(\mathbf{x})| \\ &= \sup_{\mathbf{x} \in \mathbb{R}^d} |\mathbb{1}(\mathbf{x})f(\mathbf{x})| + \max_{|\alpha|=1} \sup_{\mathbf{x} \in \mathbb{R}^d} |\mathbb{1}(\mathbf{x})\partial^\alpha f(\mathbf{x}) + f(\mathbf{x})\partial^\alpha \mathbb{1}(\mathbf{x})| \\ &\leq \sup_{\mathbf{x} \in \mathbb{R}^d} |\mathbb{1}(\mathbf{x})f(\mathbf{x})| + \max_{|\alpha|=1} \sup_{\mathbf{x} \in \mathbb{R}^d} |\mathbb{1}(\mathbf{x})\partial^\alpha f(\mathbf{x})| + \max_{|\alpha|=1} \sup_{\mathbf{x} \in \mathbb{R}^d} |f(\mathbf{x})\partial^\alpha \mathbb{1}(\mathbf{x})|. \end{aligned}$$

Recall that the infinitely differentiable function  $\mathbb{1}$  has the property  $\mathbb{1}(\mathbf{x}) = 1$  if  $\mathbf{x} \in \mathcal{X}$  and  $\mathbb{1}(\mathbf{x}) = 0$  if  $\mathbf{x} \notin [-S, S]^d$ , meaning that  $\mathbb{1}(\mathbf{x}) = 0$  and  $\partial^\alpha \mathbb{1}(\mathbf{x}) = 0$  for all  $\mathbf{x} \notin [-S, S]^d$ . By the assumption  $f \in C^1(\mathbb{R}^d)$ , the all terms  $\mathbb{1}(\mathbf{x})f(\mathbf{x})$ ,  $\mathbb{1}(\mathbf{x})\partial^\alpha f(\mathbf{x})$ , and  $f(\mathbf{x})\partial^\alpha \mathbb{1}(\mathbf{x})$  vanish outside of  $\mathbf{x} \in [-S, S]^d$ . Therefore the following inequality holds:

$$\begin{aligned} M_1(f_{\text{clip}}) &\leq \sup_{\mathbf{x} \in [-S, S]^d} |\mathbb{1}(\mathbf{x})f(\mathbf{x})| + \max_{|\alpha|=1} \sup_{\mathbf{x} \in [-S, S]^d} |\mathbb{1}(\mathbf{x})\partial^\alpha f(\mathbf{x})| + \max_{|\alpha|=1} \sup_{\mathbf{x} \in [-S, S]^d} |f(\mathbf{x})\partial^\alpha \mathbb{1}(\mathbf{x})|. \\ &\leq \sup_{\mathbf{x} \in [-S, S]^d} |\mathbb{1}(\mathbf{x})| \times \sup_{\mathbf{x} \in [-S, S]^d} |f(\mathbf{x})| + \\ &\quad \sup_{\mathbf{x} \in [-S, S]^d} |\mathbb{1}(\mathbf{x})| \times \max_{|\alpha|=1} \sup_{\mathbf{x} \in [-S, S]^d} |\partial^\alpha f(\mathbf{x})| + \sup_{\mathbf{x} \in [-S, S]^d} |f(\mathbf{x})| \times \max_{|\alpha|=1} \sup_{\mathbf{x} \in [-S, S]^d} |\partial^\alpha \mathbb{1}(\mathbf{x})| \\ &= 3M_1^*(\mathbb{1})M_1^*(f) < \infty. \end{aligned} \tag{37}$$

The quantity  $B_1(f_{\text{clip}})$  is clearly bounded since  $f_{\text{clip}}(\mathbf{x})$  is compactly support on  $[-S, S]^d$ :

$$\begin{aligned} B_1(f_{\text{clip}}) &= \int_{[-S, S]^d} f_{\text{clip}}(\mathbf{x})(1 + \|\mathbf{x}\|_2) d\mathbf{x} \\ &\leq \int_{[-S, S]^d} M_1(f_{\text{clip}}) \cdot (1 + S) d\mathbf{x} = (1 + S)(2S)^d M_1(f_{\text{clip}}) < \infty. \end{aligned} \tag{38}$$

Thus we may apply Theorem 2 to obtain

$$(*) \leq C'_1 \max(M_1(f_{\text{clip}}), B_1(f_{\text{clip}})) \left( \frac{1}{\sigma_w} + \frac{\sigma_w^d (\sigma_w + 1)}{\sigma_b} \right)$$

for some constant  $C'_1 > 0$  depending only on  $\phi, \psi, p_w, p_b$ . Let  $C_1 := (1 + S)(2S)^d C'_1$ , so that from (38) we have

$$(*) \leq C_1 M_1(f_{\text{clip}}) \left( \frac{1}{\sigma_w} + \frac{\sigma_w^d (\sigma_w + 1)}{\sigma_b} \right).$$

**Bounding (\*\*)**: Let

$$h_{\mathbf{w}, b}(\mathbf{x}) := f_{\text{clip}}(\mathbf{x}) \frac{\psi(\mathbf{w} \cdot \mathbf{x} + b)}{1 + \|\mathbf{w}\|_2}.$$

Since  $f_{\text{clip}}$  vanishes outside  $[-S, S]^d$ , so does  $h_{\mathbf{w},b}$  and therefore

$$I_\sigma f(\mathbf{x}) - I_{\sigma,D} f(\mathbf{x}) = \int_{\mathbb{R}} \int_{\mathbb{R}^d} \underbrace{\left\{ \int_{[-S,S]^d} h_{\mathbf{w},b}(\mathbf{x}') d\mathbf{x}' - \frac{2^d}{D} \sum_{j=1}^D h_{\mathbf{w},b}(\mathbf{x}_j) \right\}}_{(a)} \times \\ (1 + \|\mathbf{w}\|_2) \phi(\mathbf{w} \cdot \mathbf{x} + b) Z p_{\mathbf{w},\sigma}(\mathbf{w}) p_{b,\sigma}(b) d\mathbf{w} db.$$

Our aim here to show that the collection of functions  $h_{\mathbf{w},b}$  indexed by  $\mathbf{w}$  and  $b$  has derivatives that are uniformly bounded on  $[-S, S]^d$ , in order that Lemma 4 can be applied to (a). Let  $\boldsymbol{\alpha}$  be a multi-index such that  $|\boldsymbol{\alpha}| = 1$ . By the chain rule of differentiation,

$$\begin{aligned} \partial^\alpha h_{\mathbf{w},b}(\mathbf{x}) &= \left( \partial^\alpha f_{\text{clip}}(\mathbf{x}) \psi(\mathbf{w} \cdot \mathbf{x} + b) + f_{\text{clip}}(\mathbf{x}) \partial^\alpha \psi(\mathbf{w} \cdot \mathbf{x} + b) \right) \frac{1}{1 + \|\mathbf{w}\|_2} \\ &= \left( \partial^\alpha f_{\text{clip}}(\mathbf{x}) \psi(\mathbf{w} \cdot \mathbf{x} + b) + f_{\text{clip}}(\mathbf{x}) \mathbf{w}^\alpha (\partial \psi)(\mathbf{w} \cdot \mathbf{x} + b) \right) \frac{1}{1 + \|\mathbf{w}\|_2} \end{aligned}$$

where we recall that  $\mathbf{w}^\alpha = w_i$  for the non zero element index  $i$  of  $\boldsymbol{\alpha}$  and that  $\partial \psi$  is the first derivative of  $\psi : \mathbb{R} \rightarrow \mathbb{R}$ . Since  $|\boldsymbol{\alpha}| = 1$ , we have  $\mathbf{w}^\alpha \leq \|\mathbf{w}\|_2$ . In addition,  $\partial^\alpha f_{\text{clip}}(\mathbf{x}) \leq M_1(f_{\text{clip}})$  and  $f_{\text{clip}}(\mathbf{x}) \leq M_1(f_{\text{clip}})$  for all  $\mathbf{x} \in [-S, S]^d$  by definition. Therefore

$$\partial^\alpha h_{\mathbf{w},b}(\mathbf{x}) \leq M_1(f_{\text{clip}}) \left( \frac{1}{1 + \|\mathbf{w}\|_2} \psi(\mathbf{w} \cdot \mathbf{x} + b) + \frac{\|\mathbf{w}\|_2}{1 + \|\mathbf{w}\|_2} (\partial \psi)(\mathbf{w} \cdot \mathbf{x} + b) \right).$$

Since  $\psi \in \mathcal{S}(\mathbb{R})$ ,  $\frac{1}{1 + \|\mathbf{w}\|_2} \leq 1$ , and  $\frac{\|\mathbf{w}\|_2}{1 + \|\mathbf{w}\|_2} \leq 1$  for all  $\mathbf{w} \in \mathbb{R}^d$ , we further have

$$\partial^\alpha h_{\mathbf{w},b}(\mathbf{x}) \leq 2M_1(f_{\text{clip}}) M_1(\psi),$$

which is a uniform bound, independent of  $\mathbf{x} \in [-S, S]^d$ ,  $\mathbf{w} \in \mathbb{R}^d$ ,  $b \in \mathbb{R}$ , and  $\boldsymbol{\alpha}$  such that  $|\boldsymbol{\alpha}| = 1$ . Applying Lemma 4 for the term (a) with the grid points  $(\mathbf{x}_i)_{i=1}^D$  on  $[-S, S]^d$ ,

$$\begin{aligned} \left| \int_{[-S,S]^d} h_{\mathbf{w},b}(\mathbf{x}') d\mathbf{x}' - \frac{2^d}{D} \sum_{j=1}^D h_{\mathbf{w},b}(\mathbf{x}_j) \right| &\leq \frac{C'_Q}{D^{\frac{1}{d}}} \max_{|\boldsymbol{\alpha}|=1} \sup_{\mathbf{x}' \in [-S,S]^d} |\partial^\alpha h_{\mathbf{w},b}(\mathbf{x}')| \\ &\leq \frac{2C'_Q M_1(f_{\text{clip}}) M_1(\psi)}{D^{\frac{1}{d}}}. \end{aligned}$$

where  $C'_Q$  is a constant independent of  $h_{\mathbf{w},b}$ . Therefore setting  $C_Q = 2C'_Q$  we have that

$$\begin{aligned} |I_\sigma f(\mathbf{x}) - I_{\sigma,D} f(\mathbf{x})| &\leq \frac{C_Q M_1(f_{\text{clip}}) M_1(\psi)}{D^{\frac{1}{d}}} \underbrace{\int_{\mathbb{R}} \int_{\mathbb{R}^d} \left| (1 + \|\mathbf{w}\|_2) \phi(\mathbf{w} \cdot \mathbf{x} + b) Z p_{\mathbf{w},\sigma}(\mathbf{w}) p_{b,\sigma}(b) \right| d\mathbf{w} db}_{(b)}, \quad (39) \end{aligned}$$

where we move  $C_Q M_1(\psi) M_1(f_{\text{clip}})$  outside the integral as they are  $\mathbf{w}$  and  $b$ -independent.

It remains to bound (b). By the assumption  $\phi \in C_0^*(\mathbb{R})$ ,

$$(b) \leq Z \|\phi\|_{L^\infty} \int_{\mathbb{R}^d} (1 + \|\mathbf{w}\|_2) p_{\mathbf{w},\sigma}(\mathbf{w}) d\mathbf{w} \int_{\mathbb{R}} p_{b,\sigma}(b) db = Z \|\phi\|_{L^\infty} \int_{\mathbb{R}^d} (1 + \|\mathbf{w}\|_2) p_{\mathbf{w},\sigma}(\mathbf{w}) d\mathbf{w},$$

where  $\int_{\mathbb{R}} p_{b,\sigma}(b)db = 1$  since  $p_{b,\sigma}$  is probability density. Since  $p_{\mathbf{w},\sigma}(\mathbf{w}) = \frac{1}{\sigma_{\mathbf{w}}^d} p_{\mathbf{w}}\left(\frac{\mathbf{w}}{\sigma_{\mathbf{w}}}\right)$ , by a change of variable  $\mathbf{w}' = \sigma_{\mathbf{w}}^{-1}\mathbf{w}$ ,

$$(b) \leq Z \|\phi\|_{L^\infty} \int_{\mathbb{R}^d} (1 + \sigma_{\mathbf{w}} \|\mathbf{w}'\|_2) p_{\mathbf{w}}(\mathbf{w}') d\mathbf{w}'.$$

Since  $p_{\mathbf{w}}$  has the finite second moment, let  $C_{p_{\mathbf{w}}} := \max\left(1, \int_{\mathbb{R}^d} \|\mathbf{w}\|_2^2 p_{\mathbf{w}}(\mathbf{w}) d\mathbf{w}\right)$  to see

$$(b) \leq Z \|\phi\|_{L^\infty} C_{p_{\mathbf{w}}} (1 + \sigma_{\mathbf{w}}).$$

Recalling  $Z = (2\pi)^{\frac{1}{2}} \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \|\widehat{p_b}\|_{L^1}^{-1} \sigma_{\mathbf{w}}^d \sigma_b$ ,

$$(b) \leq (2\pi)^{\frac{1}{2}} \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \|\widehat{p_b}\|_{L^1}^{-1} \|\phi\|_{L^\infty} C_{p_{\mathbf{w}}} \sigma_b \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1).$$

By plugging the upper bound of (b) in and setting  $C_2 := (2\pi)^{\frac{1}{2}} \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \|\widehat{p_b}\|_{L^1}^{-1} \|\phi\|_{L^\infty} M_1(\psi) C_{p_{\mathbf{w}}} C_Q$ , we arrive at the overall bound

$$\sup_{\mathbf{x} \in \mathcal{X}} |I_\sigma f(\mathbf{x}) - I_{\sigma,D} f(\mathbf{x})| \leq C_2 M_1(f_{\text{clip}}) \frac{\sigma_b \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)}{D^{\frac{1}{d}} (\log D)^{-1}}.$$

**Bounding (\*\*\*)**: Define  $\tau_\phi(\mathbf{x}, \mathbf{w}, b) := \left( \frac{(2S)^d}{D} \sum_{j=1}^D f_{\text{clip}}(\mathbf{x}'_j) \psi(\mathbf{w} \cdot \mathbf{x}'_j + b) \right) \phi(\mathbf{w} \cdot \mathbf{x} + b)$ .

Then the term (\*\*\*) can be written as

$$\sup_{\mathbf{x} \in \mathcal{X}} |I_{\sigma,D} f(\mathbf{x}) - I_{\sigma,D,N} f(\mathbf{x})| = Z \sup_{\mathbf{x} \in \mathcal{X}} \left| \mathbb{E}_{(\mathbf{w},b)} [\tau_\phi(\mathbf{x}, \mathbf{w}, b)] - \frac{1}{N} \sum_{i=1}^N \tau_\phi(\mathbf{x}, \mathbf{w}_i, b_i) \right|.$$

We apply Lemma 2 and Lemma 3 to obtain the upper bound of (\*\*\*). In order to apply Lemma 2, it is to be verified that there exists  $G : \mathbb{R}^{d+1} \rightarrow \mathbb{R}$  such that  $\mathbb{E}_{(\mathbf{w},b)} [G(\mathbf{w}, b)^2] < \infty$  and  $|\tau_\phi(\mathbf{x}, \mathbf{w}, b) - \tau_\phi(\mathbf{x}', \mathbf{w}, b)| \leq G(\mathbf{w}, b) \|\mathbf{x} - \mathbf{x}'\|_2$ .

Recalling that  $\|f_{\text{clip}}\|_{L^\infty} \leq M_1(f_{\text{clip}})$  and that  $\phi$  was assumed to be Lipschitz with constant denoted  $L_\phi$ , the difference of  $\tau_\phi$  is given by

$$\begin{aligned} |\tau_\phi(\mathbf{x}, \mathbf{w}, b) - \tau_\phi(\mathbf{x}', \mathbf{w}, b)| &= \left| \left( \frac{(2S)^d}{D} \sum_{j=1}^D f_{\text{clip}}(\mathbf{x}'_j) \psi(\mathbf{w} \cdot \mathbf{x}'_j + b) \right) (\phi(\mathbf{w} \cdot \mathbf{x} + b) - \phi(\mathbf{w} \cdot \mathbf{x}' + b)) \right| \\ &\leq (2S)^d M_1(f_{\text{clip}}) \|\psi\|_{L^\infty} L_\phi |\mathbf{w} \cdot (\mathbf{x} - \mathbf{x}')| \\ &\leq (2S)^d M_1(f_{\text{clip}}) \|\psi\|_{L^\infty} L_\phi \|\mathbf{w}\|_2 \|\mathbf{x} - \mathbf{x}'\|_2 \end{aligned}$$

where the final inequality used Cauchy-Schwartz. Let  $G(\mathbf{w}, b) := M_1(f_{\text{clip}}) \|\psi\|_{L^\infty} L_\phi \|\mathbf{w}\|_2$ , so that

$$\begin{aligned} \mathbb{E}_{(\mathbf{w},b)} [G(\mathbf{w}, b)^2] &= (2S)^{2d} M_1(f_{\text{clip}})^2 \|\psi\|_{L^\infty}^2 L_\phi^2 \int_{\mathbb{R}^d} \|\mathbf{w}\|_2^2 p_{\mathbf{w},\sigma}(\mathbf{w}) d\mathbf{w} \int_{\mathbb{R}} p_{b,\sigma}(b) db \\ &= (2S)^{2d} M_1(f_{\text{clip}})^2 \|\psi\|_{L^\infty}^2 L_\phi^2 \int_{\mathbb{R}^d} \|\mathbf{w}\|_2^2 p_{\mathbf{w},\sigma}(\mathbf{w}) d\mathbf{w}. \end{aligned}$$



Since  $p_{\mathbf{w},\sigma}(\mathbf{w}) = \frac{1}{\sigma_{\mathbf{w}}^d} p_{\mathbf{w}}\left(\frac{\mathbf{w}}{\sigma_{\mathbf{w}}}\right)$ , by a change of variable  $\mathbf{w}' = \sigma_{\mathbf{w}}^{-1}\mathbf{w}$ , we have  $\int_{\mathbb{R}^d} \|\mathbf{w}\|_2^2 p_{\mathbf{w},\sigma}(\mathbf{w}) d\mathbf{w} = \sigma_{\mathbf{w}}^2 \int_{\mathbb{R}^d} \|\mathbf{w}'\|_2^2 p_{\mathbf{w}}(\mathbf{w}') d\mathbf{w}'$ . By the assumption that  $p_{\mathbf{w}}$  has the finite second moment, let  $V_{p_{\mathbf{w}}}^2 := \int_{\mathbb{R}^d} \|\mathbf{w}\|_2^2 p_{\mathbf{w}}(\mathbf{w}) d\mathbf{w}$  to see

$$\mathbb{E}_{(\mathbf{w},b)} [G(\mathbf{w}, b)^2] \leq \left\{ (2S)^d M_1(f_{\text{clip}}) \|\psi\|_{L^\infty} L_\phi V_{p_{\mathbf{w}}} \sigma_{\mathbf{w}} \right\}^2.$$

By Lemma 2, for some constant  $C_{\mathcal{X}}$  only depending on  $\mathcal{X}$ ,

$$\begin{aligned} \mathbb{E} \left[ \sup_{\mathbf{x} \in \mathcal{X}} \left| \mathbb{E}_{(\mathbf{w},b)} [\tau_\phi(\mathbf{x}, \mathbf{w}, b)] - \frac{1}{N} \sum_{i=1}^N \tau_\phi(\mathbf{x}, \mathbf{w}_i, b_i) \right| \right] &\leq \frac{C_{\mathcal{X}} \sqrt{\mathbb{E}_{(\mathbf{w},b)} [G(\mathbf{w}, b)^2]}}{\sqrt{N}} \\ &\leq \frac{C_{\mathcal{X}} (2S)^d M_1(f_{\text{clip}}) \|\psi\|_{L^\infty} L_\phi V_{p_{\mathbf{w}}} \sigma_{\mathbf{w}}}{\sqrt{N}}. \end{aligned} \quad (40)$$

By the upper bound  $\left( \frac{(2S)^d}{D} \sum_{j=1}^D f_{\text{clip}}(\mathbf{x}'_j) \psi(\mathbf{w} \cdot \mathbf{x}'_j + b) \right) \leq (2S)^d M_1(f_{\text{clip}}) \|\psi\|_{L^\infty}$  and the assumption  $\phi \in C_0^*(\mathbb{R})$ , we have  $\tau_\phi(\mathbf{x}, \mathbf{w}, b) \leq (2S)^d M_1(f_{\text{clip}}) \|\psi\|_{L^\infty} \|\phi\|_{L^\infty}$  for all  $(\mathbf{x}, \mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ . From Lemma 3, we have, with probability at least  $1 - \delta$ ,

$$\sup_{\mathbf{x} \in \mathcal{X}} |I_{\sigma,D} f(\mathbf{x}) - I_{\sigma,D,N} f(\mathbf{x})| \leq Z (2S)^d M_1(f_{\text{clip}}) \|\psi\|_{L^\infty} \left( \frac{C_{\mathcal{X}} L_\phi V_{p_{\mathbf{w}}} \sigma_{\mathbf{w}}}{\sqrt{N}} + \frac{\sqrt{2} \|\phi\|_{L^\infty} \sqrt{\log \delta^{-1}}}{\sqrt{N}} \right).$$

Let  $C_3 := (2\pi)^{\frac{1}{2}} (2S)^d \|\widehat{p}_{\mathbf{w}}\|_{L^1}^{-1} \|\widehat{p}_b\|_{L^1}^{-1} \|\psi\|_{L^\infty} \max(C_{\mathcal{X}} L_\phi V_{p_{\mathbf{w}}}, \sqrt{2} \|\phi\|_{L^\infty})$  where we recall that  $Z = (2\pi)^{\frac{1}{2}} \|\widehat{p}_{\mathbf{w}}\|_{L^1}^{-1} \|\widehat{p}_b\|_{L^1}^{-1} \sigma_{\mathbf{w}}^d \sigma_b$ . Then we have

$$\sup_{\mathbf{x} \in \mathcal{X}} |I_{\sigma,D} f(\mathbf{x}) - I_{\sigma,D,N} f(\mathbf{x})| \leq C_3 M_1(f_{\text{clip}}) \left( \frac{\sigma_b \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + \sqrt{\log \delta^{-1}})}{\sqrt{N}} \right).$$

For all bounds of (\*), (\*\*) and (\*\*\*) , recall that  $M(f_{\text{clip}}) \leq 3M_1^*(\mathbb{1})M_1^*(f)$  from (37). Then combining (\*), (\*\*) and (\*\*\*) and setting  $C := 3M_1^*(\mathbb{1})(C_1 + C_2 + C_3)$  completes the proof.  $\blacksquare$

### A.2.3 PROOF OF THEOREM 1

**Proof** For a bivariate function  $g(\mathbf{x}, \mathbf{y})$  we let  $I_{\sigma,D,N}^{\mathbf{x}} g(\mathbf{x}, \mathbf{y})$  denote the action of  $I_{\sigma,D,N}$  on the first argument of  $g$  and we let  $I_{\sigma,D,N}^{\mathbf{y}} g(\mathbf{x}, \mathbf{y})$  denote the action of  $I_{\sigma,D,N}$  on the second argument of  $g$ . To reduce notation, in this proof we denote  $\mathbb{E}_{f|\mathbf{w},b}[\cdot] := \mathbb{E}[\cdot \mid \{\mathbf{w}_i, b_i\}_{i=1}^N]$ . For fixed  $\mathbf{x}, \mathbf{y}$  we let

$$\begin{aligned} (a) &:= \mathbb{E}_{f|\mathbf{w},b} [(f(\mathbf{x}) - I_{\sigma,D,N} f(\mathbf{x})) (f(\mathbf{y}) - I_{\sigma,D,N} f(\mathbf{y}))] \\ &= \mathbb{E}_{f|\mathbf{w},b} [f(\mathbf{x}) f(\mathbf{y})] - \mathbb{E}_{f|\mathbf{w},b} [I_{\sigma,D,N} f(\mathbf{x}) f(\mathbf{y})] - \mathbb{E}_{f|\mathbf{w},b} [f(\mathbf{x}) I_{\sigma,D,N} f(\mathbf{y})] \\ &\quad + \mathbb{E}_{f|\mathbf{w},b} [I_{\sigma,D,N} f(\mathbf{x}) I_{\sigma,D,N} f(\mathbf{y})]. \end{aligned}$$

Recall that  $\mathbb{E}_{f|\mathbf{w},b}[f(\mathbf{x})f(\mathbf{y})] = k(\mathbf{x}, \mathbf{y}) + m(\mathbf{x})m(\mathbf{y})$  for  $f \sim \mathcal{GP}(m, k)$ . Then

$$\begin{aligned}
 (a) &= \mathbb{E}_{f|\mathbf{w},b}[f(\mathbf{x})f(\mathbf{y})] - I_{\sigma,D,N}^{\mathbf{x}}\mathbb{E}_{f|\mathbf{w},b}[f(\mathbf{x})f(\mathbf{y})] - I_{\sigma,D,N}^{\mathbf{y}}\mathbb{E}_{f|\mathbf{w},b}[f(\mathbf{x})f(\mathbf{y})] \\
 &\quad + I_{\sigma,D,N}^{\mathbf{x}}I_{\sigma,D,N}^{\mathbf{y}}\mathbb{E}_{f|\mathbf{w},b}[f(\mathbf{x})f(\mathbf{y})] \\
 &= (k(\mathbf{x}, \mathbf{y}) + m(\mathbf{x})m(\mathbf{y})) - I_{\sigma,D,N}^{\mathbf{x}}(k(\mathbf{x}, \mathbf{y}) + m(\mathbf{x})m(\mathbf{y})) \\
 &\quad - I_{\sigma,D,N}^{\mathbf{y}}(k(\mathbf{x}, \mathbf{y}) + m(\mathbf{x})m(\mathbf{y})) + I_{\sigma,D,N}^{\mathbf{x}}I_{\sigma,D,N}^{\mathbf{y}}(k(\mathbf{x}, \mathbf{y}) + m(\mathbf{x})m(\mathbf{y})) \\
 &= (m(\mathbf{x})m(\mathbf{y}) - I_{\sigma,D,N}m(\mathbf{x})m(\mathbf{y}) - m(\mathbf{x})I_{\sigma,D,N}m(\mathbf{y}) + I_{\sigma,D,N}m(\mathbf{x})I_{\sigma,D,N}m(\mathbf{y})) \\
 &\quad + \left( k(\mathbf{x}, \mathbf{y}) - I_{\sigma,D,N}^{\mathbf{x}}k(\mathbf{x}, \mathbf{y}) - I_{\sigma,D,N}^{\mathbf{y}}k(\mathbf{x}, \mathbf{y}) + I_{\sigma,D,N}^{\mathbf{x}}I_{\sigma,D,N}^{\mathbf{y}}k(\mathbf{x}, \mathbf{y}) \right)
 \end{aligned}$$

Let  $h(\mathbf{x}, \mathbf{y}) := k(\mathbf{x}, \mathbf{y}) - I_{\sigma,D,N}^{\mathbf{x}}k(\mathbf{x}, \mathbf{y})$  in order to see

$$(a) = (m(\mathbf{x}) - I_{\sigma,D,N}m(\mathbf{x}))(m(\mathbf{y}) - I_{\sigma,D,N}m(\mathbf{y})) + \left( h(\mathbf{x}, \mathbf{y}) - I_{\sigma,D,N}^{\mathbf{y}}h(\mathbf{x}, \mathbf{y}) \right).$$

Therefore the error is

$$\begin{aligned}
 \sup_{\mathbf{x} \in \mathcal{X}} \sqrt{\mathbb{E}_{f|\mathbf{w},b} \left[ (f(\mathbf{x}) - I_{\sigma,D,N}f(\mathbf{x}))^2 \right]} &\leq \sup_{\mathbf{x} \in [-S,S]^d} \sqrt{\mathbb{E}_{f|\mathbf{w},b} \left[ (f(\mathbf{x}) - I_{\sigma,D,N}f(\mathbf{x}))^2 \right]} \\
 &\leq \underbrace{\sup_{\mathbf{x} \in [-S,S]^d} |m(\mathbf{x}) - I_{\sigma,D,N}m(\mathbf{x})|}_{(*)} + \underbrace{\sup_{\mathbf{x} \in [-S,S]^d} \left| h(\mathbf{x}, \mathbf{x}) - I_{\sigma,D,N}^{\mathbf{y}}h(\mathbf{x}, \mathbf{x}) \right|}_{(**)}^{\frac{1}{2}}
 \end{aligned}$$

In the remainder we bound  $(*)$  and  $(**)$ .

**Bounding  $(*)$ :** Applying Theorem 3, we immediately have, with probability at least  $1 - \delta$  with respect to the random variables  $\{\mathbf{w}_i, b_i\}_{i=1}^N$ ,

$$(*) \leq C_1 M_1^*(m) \left( \frac{1}{\sigma_{\mathbf{w}}} + \frac{\sigma_{\mathbf{w}}^d(\sigma_{\mathbf{w}} + 1)}{\sigma_b} + \frac{\sigma_b \sigma_{\mathbf{w}}^d(\sigma_{\mathbf{w}} + 1)}{D^{\frac{1}{d}}} + \frac{\sigma_b \sigma_{\mathbf{w}}^d(\sigma_{\mathbf{w}} + \sqrt{\log \delta^{-1}})}{\sqrt{N}} \right)$$

for some constant  $C_1$  independent of  $\delta, \sigma_{\mathbf{x}}, \sigma_b, D, N$  and  $m$ .

**Bounding  $(**)$ :** It is clear to see

$$(**) = \left\{ \sup_{\mathbf{x} \in [-S,S]^d} \left| h(\mathbf{x}, \mathbf{x}) - I_{\sigma,D,N}^{\mathbf{y}}h(\mathbf{x}, \mathbf{x}) \right| \right\}^{\frac{1}{2}} \leq \left\{ \sup_{\mathbf{x} \in [-S,S]^d} \sup_{\mathbf{y} \in [-S,S]^d} \left| h(\mathbf{x}, \mathbf{y}) - I_{\sigma,D,N}^{\mathbf{y}}h(\mathbf{x}, \mathbf{y}) \right| \right\}^{\frac{1}{2}} \quad (41)$$

First, with respect to the supremum of  $\mathbf{y}$ , from Theorem 3, with probability at least  $1 - \delta$  with respect to the random variables  $\{\mathbf{w}_i, b_i\}_{i=1}^N$ ,

$$\begin{aligned}
 &\sup_{\mathbf{y} \in [-S,S]^d} \left| h(\mathbf{x}, \mathbf{y}) - I_{\sigma,D,N}^{\mathbf{y}}h(\mathbf{x}, \mathbf{y}) \right| \\
 &\leq C_2' M_1^*(h(\mathbf{x}, \cdot)) \left( \frac{1}{\sigma_{\mathbf{w}}} + \frac{\sigma_{\mathbf{w}}^d(\sigma_{\mathbf{w}} + 1)}{\sigma_b} + \frac{\sigma_b \sigma_{\mathbf{w}}^d(\sigma_{\mathbf{w}} + 1)}{D^{\frac{1}{d}}} + \frac{\sigma_b \sigma_{\mathbf{w}}^d(\sigma_{\mathbf{w}} + \sqrt{\log \delta^{-1}})}{\sqrt{N}} \right) \quad (42)
 \end{aligned}$$

where  $M_1^*(h(\mathbf{x}, \cdot))$  is given as

$$M_1^*(h(\mathbf{x}, \cdot)) = \max_{|\beta| \leq 1} \sup_{\mathbf{y} \in [-S, S]^d} |\partial^{0, \beta} h(\mathbf{x}, \mathbf{y})|. \quad (43)$$

Second,  $|M_1^*(h(\mathbf{x}, \cdot))|$  is to be bounded. Recall  $h(\mathbf{x}, \mathbf{y}) = k(\mathbf{x}, \mathbf{y}) - I_{\sigma, D, N}^{\mathbf{x}} k(\mathbf{x}, \mathbf{y})$  to see  $\partial^{0, \beta} h(\mathbf{x}, \mathbf{y}) = \partial^{0, \beta} k(\mathbf{x}, \mathbf{y}) - I_{\sigma, D, N}^{\mathbf{x}} \partial^{0, \beta} k(\mathbf{x}, \mathbf{y})$ . For fixed  $\mathbf{y} \in [-S, S]^d$  and  $|\beta| \leq 1$ ,  $\partial^{0, \beta} h(\mathbf{x}, \mathbf{y})$  is upper bounded for all  $\mathbf{x} \in [-S, S]^d$  from Theorem 3: with probability at least  $1 - \delta$  with respect to the random variables  $\{\mathbf{w}_i, b_i\}_{i=1}^N$ ,

$$\begin{aligned} \partial^{0, \beta} h(\mathbf{x}, \mathbf{y}) &\leq \sup_{\mathbf{x} \in [-S, S]^d} \left| \partial^{0, \beta} h(\mathbf{x}, \mathbf{y}) \right| \\ &= \sup_{\mathbf{x} \in [-S, S]^d} \left| \partial^{0, \beta} k(\mathbf{x}, \mathbf{y}) - I_{\sigma, D, N}^{\mathbf{x}} \partial^{0, \beta} k(\mathbf{x}, \mathbf{y}) \right| \\ &\leq C_2'' M_1^*(\partial^{0, \beta} k(\cdot, \mathbf{y})) \left( \frac{1}{\sigma_{\mathbf{w}}} + \frac{\sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)}{\sigma_b} + \frac{\sigma_b \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)}{D^{\frac{1}{d}}} + \frac{\sigma_b \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + \sqrt{\log \delta^{-1}})}{\sqrt{N}} \right). \end{aligned}$$

where  $M_1^*(\partial^{0, \beta} k(\cdot, \mathbf{y})) = \max_{|\alpha| \leq 1} \sup_{\mathbf{x} \in [-S, S]^d} |\partial^{\alpha, \beta} k(\mathbf{x}, \mathbf{y})|$ . Plugging this upper bound into (43), with probability at least  $1 - \delta$  with respect to the random variables  $\{\mathbf{w}_i, b_i\}_{i=1}^N$ ,

$$\begin{aligned} M_1^*(h(\mathbf{x}, \cdot)) &\leq C_2'' \max_{|\beta| \leq 1} \sup_{\mathbf{y} \in [-S, S]^d} \left| \max_{|\alpha| \leq 1} \sup_{\mathbf{x} \in [-S, S]^d} \left| \partial^{\alpha, \beta} k(\mathbf{x}, \mathbf{y}) \right| \right| \\ &\quad \left( \frac{1}{\sigma_{\mathbf{w}}} + \frac{\sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)}{\sigma_b} + \frac{\sigma_b \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)}{D^{\frac{1}{d}}} + \frac{\sigma_b \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + \sqrt{\log \delta^{-1}})}{\sqrt{N}} \right). \end{aligned}$$

Recall  $M_1^*(k) := \max_{|\alpha| \leq 1, |\beta| \leq 1} \sup_{\mathbf{x}, \mathbf{y} \in [-S, S]^d} |\partial^{\alpha, \beta} k(\mathbf{x}, \mathbf{y})|$  to see, for all  $\mathbf{x} \in [-S, S]^d$ , and with probability at least  $1 - \delta$  with respect to the random variables  $\{\mathbf{w}_i, b_i\}_{i=1}^N$ ,

$$M_1^*(h(\mathbf{x}, \cdot)) \leq C_2'' M_1^*(k) \left( \frac{1}{\sigma_{\mathbf{w}}} + \frac{\sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)}{\sigma_b} + \frac{\sigma_b \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)}{D^{\frac{1}{d}}} + \frac{\sigma_b \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + \sqrt{\log \delta^{-1}})}{\sqrt{N}} \right).$$

Notice that  $M_1^*(k) < \infty$  by the assumption  $k \in C^{1 \times 1}(\mathbb{R}^d \times \mathbb{R}^d)$ . Combining this upper bound with (42) and taking the supremum over  $\mathbf{x} \in [-S, S]^d$ , from (41) we have, with probability at least  $1 - \delta$  with respect to the random variables  $\{\mathbf{w}_i, b_i\}_{i=1}^N$ ,

$$(**) \leq C_2 \sqrt{M_1^*(k)} \left( \frac{1}{\sigma_{\mathbf{w}}} + \frac{\sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)}{\sigma_b} + \frac{\sigma_b \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)}{D^{\frac{1}{d}}} + \frac{\sigma_b \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + \sqrt{\log \delta^{-1}})}{\sqrt{N}} \right).$$

where  $C_2 := \sqrt{C_2' C_2''}$ .

Combining these bounds on (\*) and (\*\*) completes the proof.  $\blacksquare$

### A.3 An Analogous Result to Theorem 1 for Unbounded $\phi$

In this section we state and prove an analogous result of Theorem 1 that holds under weaker assumptions on the activation function  $\phi$ , with a correspondingly stronger assumption on the GP. This ensures that our theory is compatible with activation functions  $\phi$  that may be unbounded, including the ReLU activation function  $\phi(x) = \max(0, x)$ . An example of the associated pair function  $\psi$  is given in Table 1.

First of all, we recall that a positive semi-definite function  $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  reproduces a Hilbert space  $\mathcal{H}_k$  whose elements are functions  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  such that  $k(\cdot, \mathbf{x}) \in \mathcal{H}_k$  for all  $\mathbf{x} \in \mathbb{R}^d$  and  $\langle f, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_k} = f(\mathbf{x})$  for all  $\mathbf{x} \in \mathbb{R}^d$  and all  $f \in \mathcal{H}_k$ . The Hilbert space  $\mathcal{H}_k$  is called a *reproducing kernel Hilbert space* (RKHS). It is a well-known fact that GP sample paths are not contained in the RKHS induced by the GP covariance kernel i.e.  $f \sim \mathcal{GP}(m, k) \implies \mathbb{P}(f \in \mathcal{H}_k) = 0$ , unless  $\mathcal{H}_k$  is finite dimensional (Hsing and Eubank, 2015, Theorem 7.5.4). However,  $\mathbb{P}(f \in \mathcal{H}_R) = 1$  holds whenever  $\mathcal{H}_R$  satisfies *nuclear dominance* over  $\mathcal{H}_k$ ; see Lukić and Beder (2001). The additional assumption that we require on the GP in this appendix is that the GP takes values in a RKHS  $\mathcal{H}_R$  where  $R$  is continuously differentiable. Intuitively, this imposes an additional smoothness requirement on the GP compared to Theorem 1. Our analogous result to Theorem 1 is as follows:

**Theorem 4** (Analogue of Theorem 1 for Unbounded  $\phi$ ). *In the same setting of Theorem 1, replace the assumption  $\phi \in C_0^*(\mathbb{R})$  with  $\phi \in C_1^*(\mathbb{R})$ . In addition, assume that  $f \sim \mathcal{GP}(m, k)$  is a random variable taking values in  $\mathcal{H}_R$  with the reproducing kernel  $R \in C^{1 \times 1}(\mathbb{R}^d \times \mathbb{R}^d)$ . Assume that  $m \in \mathcal{H}_R$  and that the covariance operator  $\mathcal{K}$  of  $f$  is trace class. Then, with probability at least  $1 - \delta$ ,*

$$\begin{aligned} & \sup_{\mathbf{x} \in \mathcal{X}} \sqrt{\mathbb{E} \left[ (f(\mathbf{x}) - I_{\sigma, D, N} f(\mathbf{x}))^2 \mid \{\mathbf{w}_i, b_i\}_{i=1}^N \right]} \\ & \leq C \sqrt{M_1^*(R)} \left( \|m\|_{\mathcal{H}_R} + \sqrt{\text{tr}(\mathcal{K})} \right) \left\{ \frac{1}{\sigma_{\mathbf{w}}} + \frac{\sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)^2}{\sigma_b} + \frac{\sigma_b (\sigma_b + 1) \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)^2}{D^{\frac{1}{d}}} + \frac{\sigma_b \sigma_{\mathbf{w}}^{d+1}}{\delta \sqrt{N}} \right\}. \end{aligned}$$

where  $\text{tr}(\mathcal{K})$  is the trace of the operator  $\mathcal{K}$  and  $C$  is a constant independent of  $m, k, \sigma_{\mathbf{w}}, \sigma_b, D, N, \delta$ .

**Proof** To reduce notation, in this proof we denote  $\mathbb{E}_{f|\mathbf{w}, b}[\cdot] := \mathbb{E}[\cdot \mid \{\mathbf{w}_i, b_i\}_{i=1}^N]$ . By Jensen's inequality, we have

$$\sup_{\mathbf{x} \in \mathcal{X}} \sqrt{\mathbb{E}_{f|\mathbf{w}, b} \left[ (f(\mathbf{x}) - I_{\sigma, D, N} f(\mathbf{x}))^2 \right]} \leq \sqrt{\mathbb{E}_{f|\mathbf{w}, b} \left[ \left( \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x}) - I_{\sigma, D, N} f(\mathbf{x})| \right)^2 \right]} \quad (44)$$

By Theorem 6 we have, with probability  $1 - \delta$  with respect to the random variables  $\{\mathbf{w}_i, b_i\}_{i=1}^N$ , the right hand side of (44) can be bounded as

$$\begin{aligned} & \leq \sqrt{\mathbb{E}_{f|\mathbf{w}, b} \left[ C^2 M_1^*(f)^2 \left\{ \frac{1}{\sigma_{\mathbf{w}}} + \frac{\sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)^2}{\sigma_b} + \frac{\sigma_b (\sigma_b + 1) \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)^2}{D^{\frac{1}{d}}} + \frac{\sigma_b \sigma_{\mathbf{w}}^{d+1}}{\delta \sqrt{N}} \right\}^2 \right]} \\ & = C \sqrt{\mathbb{E}_{f|\mathbf{w}, b} [M_1^*(f)^2]} \left\{ \frac{1}{\sigma_{\mathbf{w}}} + \frac{\sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)^2}{\sigma_b} + \frac{\sigma_b (\sigma_b + 1) \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)^2}{D^{\frac{1}{d}}} + \frac{\sigma_b \sigma_{\mathbf{w}}^{d+1}}{\delta \sqrt{N}} \right\}, \end{aligned}$$

where  $C$  is a constant independent on  $f, \sigma_{\mathbf{w}}, \sigma_b, \delta$ . Next we will upper-bound  $\sqrt{\mathbb{E}_{f|\mathbf{w},b} [M_1^*(f)^2]}$ . Since  $f$  is a  $\mathcal{H}_R$ -valued random variable, by the reproducing property of  $\mathcal{H}_R$

$$\mathbb{E}_{f|\mathbf{w},b} [M_1^*(f)^2] = \mathbb{E}_{f|\mathbf{w},b} \left[ \left( \max_{|\alpha| \leq 1} \sup_{\mathbf{x} \in [-S,S]^d} \partial^\alpha f(\mathbf{x}) \right)^2 \right] = \mathbb{E}_{f|\mathbf{w},b} \left[ \left( \max_{|\alpha| \leq 1} \sup_{\mathbf{x} \in [-S,S]^d} \langle f, \partial^{\alpha,0} R(\mathbf{x}, \cdot) \rangle_{\mathcal{H}_R} \right)^2 \right].$$

By the Cauchy-Schwartz inequality,

$$\begin{aligned} \mathbb{E}_{f|\mathbf{w},b} \left[ \left( \max_{|\alpha| \leq 1} \sup_{\mathbf{x} \in [-S,S]^d} \langle f, \partial^{\alpha,0} R(\mathbf{x}, \cdot) \rangle_{\mathcal{H}_R} \right)^2 \right] &\leq \mathbb{E}_{f|\mathbf{w},b} \left[ \left( \max_{|\alpha| \leq 1} \sup_{\mathbf{x} \in [-S,S]^d} \|f\|_{\mathcal{H}_R} \|\partial^{\alpha,0} R(\mathbf{x}, \cdot)\|_{\mathcal{H}_R} \right)^2 \right] \\ &= \mathbb{E}_{f|\mathbf{w},b} \left[ \max_{|\alpha| \leq 1} \sup_{\mathbf{x} \in [-S,S]^d} \|f\|_{\mathcal{H}_R}^2 \|\partial^{\alpha,0} R(\mathbf{x}, \mathbf{x})\|_{\mathcal{H}_R} \right] \\ &\leq M_1^*(R) \mathbb{E}_{f|\mathbf{w},b} [\|f\|_{\mathcal{H}_R}^2]. \end{aligned}$$

From (Prato, 2006, (1.13)),  $\mathbb{E}_{f|\mathbf{w},b} [\|f\|_{\mathcal{H}_R}^2] = \|m\|_{\mathcal{H}_R}^2 + \text{tr}(\mathcal{K})$ . Therefore, we have

$$\sqrt{\mathbb{E}_{f|\mathbf{w},b} [M_1^*(f)^2]} \leq \sqrt{M_1^*(R) (\|m\|_{\mathcal{H}_R}^2 + \text{tr}(\mathcal{K}))} \leq \sqrt{M_1^*(R)} (\|m\|_{\mathcal{H}_R} + \sqrt{\text{tr}(\mathcal{K})})$$

where the fact that  $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$  for  $a, b \in \mathbb{R}$  is applied for the last inequality. Plugging in this upper bound concludes the proof.  $\blacksquare$

In the remaining part of this section, we show Theorem 5 and Theorem 6 which are analogous results to Theorem 2 and Theorem 3 for  $\phi \in C_1^*(\mathbb{R})$ . The assumption  $\phi \in C_1^*(\mathbb{R})$  implies for some  $C_\phi < \infty$

$$\begin{aligned} \phi(\mathbf{w} \cdot \mathbf{x} + b) &\leq C_\phi (1 + |\mathbf{w} \cdot \mathbf{x} + b|) \\ &\leq C_\phi (1 + \|\mathbf{w}\|_2 \|\mathbf{x}\|_2 + |b|) \\ &\leq C_\phi (1 + \|\mathbf{x}\|_2) (1 + \|\mathbf{w}\|_2) (1 + |b|) \end{aligned} \tag{45}$$

where Cauchy-Schwartz inequality is applied for the second inequality. The same discussions in the proof of Theorem 2 and Theorem 3 holds by replacing all bounds involving  $\phi$  with an expression similar to (45). Let  $B_2(f) := \int_{\mathbb{R}^d} |f(\mathbf{x})| (1 + \|\mathbf{x}\|_2)^2 d\mathbf{x}$ .

**Theorem 5** (Analogue of Theorem 2 for Unbounded  $\phi$ ). *Let  $\mathcal{X} \subset \mathbb{R}^d$  be bounded. Let Assumption 1 and Assumption 4 hold, but with  $\phi \in C_0^*(\mathbb{R})$  replaced with  $\phi \in C_1^*(\mathbb{R})$ , and let  $f \in C^1(\mathbb{R}^d)$  satisfy  $M_1(f) < \infty$  and  $B_2(f) < \infty$ . Then*

$$\sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x}) - (R_\sigma^* R)[f](\mathbf{x})| \leq C \max(M_1(f), B_2(f)) \left\{ \frac{1}{\sigma_{\mathbf{w}}} + \frac{\sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)^2}{\sigma_b} \right\} \tag{46}$$

for some constant  $C$  that is independent of  $\sigma_{\mathbf{w}}, \sigma_b$  and  $f$ , but may depend on  $\phi, \psi, p_{\mathbf{w}}$  and  $p_b$ .

**Proof** We use the same proof as Theorem 2, but consider the supremum error over a bounded domain  $\mathcal{X} \subset \mathbb{R}^d$  i.e.  $\sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x}) - (R_\sigma^* R)[f](\mathbf{x})|$ , instead of error over  $\mathbb{R}^d$ . Recall the overall structure of the proof is (18) and in particular there are two quantities, (\*) and (\*\*) to be bounded. The argument used to bound (\*) remains valid, so our attention below turns to the argument used to bound (\*\*).

In order to establish a bound on (\*\*), we replace the upper bound involving  $\phi$  subsequent to (30). From (30), we have

$$|(3)| \leq \|\partial p_b\|_{L^\infty} (1 + \|\mathbf{x}'\|_2) (1 + \|\mathbf{w}\|_2) \int_{\mathbb{R}} (1 + |b'|) |\phi(\mathbf{w} \cdot (\mathbf{x} - \mathbf{x}') + b') \psi(b')| db'. \quad (47)$$

From (45), for some  $C_\phi < \infty$ ,

$$\begin{aligned} \phi(\mathbf{w} \cdot (\mathbf{x} - \mathbf{x}') + b') &\leq C_\phi (1 + \|\mathbf{x} - \mathbf{x}'\|_2) (1 + \|\mathbf{w}\|_2) (1 + |b|) \\ &\leq C_\phi (1 + \|\mathbf{x}\|_2) (1 + \|\mathbf{x}'\|_2) (1 + \|\mathbf{w}\|_2) (1 + |b|). \end{aligned}$$

Plugging this upper bound in (47),

$$|(3)| \leq C_\phi \|\partial p_b\|_{L^\infty} (1 + \|\mathbf{x}\|_2) (1 + \|\mathbf{x}'\|_2)^2 (1 + \|\mathbf{w}\|_2)^2 \int_{\mathbb{R}} (1 + |b'|)^2 |\psi(b')| db'. \quad (48)$$

Let  $C_\psi := \int_{\mathbb{R}} (1 + |b|)^2 |\psi(b)| db$  and  $C_{p_w} := \max(1, 2 \int_{\mathbb{R}^d} \|\mathbf{w}\|_2 p_w(\mathbf{w}) d\mathbf{w}, \int_{\mathbb{R}^d} \|\mathbf{w}\|_2^2 p_w(\mathbf{w}) d\mathbf{w})$ . By the same discussion in the proof of Theorem 2 subsequent to (30), considering the difference between (31) and (48),

$$\Delta k(\mathbf{x}, \mathbf{x}') \leq (2\pi)^{\frac{1}{2}} \|\widehat{p_w}\|_{L^1}^{-1} \|\widehat{p_b}\|_{L^1}^{-1} \|\partial p_b\|_{L^\infty} \|\phi\|_{L^\infty} C_\psi C_{p_w} (1 + \|\mathbf{x}\|_2) (1 + \|\mathbf{x}'\|_2)^2 \frac{1}{\epsilon_w^d} \left(1 + \frac{1}{\epsilon_w}\right)^2 \epsilon_b.$$

Set  $C'_2 := (2\pi)^{\frac{1}{2}} \|\widehat{p_w}\|_{L^1}^{-1} \|\widehat{p_b}\|_{L^1}^{-1} \|\partial p_b\|_{L^\infty} \|\phi\|_{L^\infty} C_\psi C_{p_w}$ . The error term (\*\*) is then bounded as

$$\begin{aligned} (**) &= \sup_{\mathbf{x} \in \mathcal{X}} \left| \int_{\mathbb{R}^d} f(\mathbf{x}') \Delta k(\mathbf{x}, \mathbf{x}') d\mathbf{x}' \right| \\ &\leq C'_2 \sup_{\mathbf{x} \in \mathcal{X}} |1 + \|\mathbf{x}\|_2| \int_{\mathbb{R}^d} (1 + \|\mathbf{x}'\|_2)^2 |f(\mathbf{x}')| d\mathbf{x}' \frac{1}{\epsilon_w^d} \left(1 + \frac{1}{\epsilon_w}\right)^2 \epsilon_b. \end{aligned}$$

Setting  $C_2 := C'_2 \sup_{\mathbf{x} \in \mathcal{X}} |1 + \|\mathbf{x}\|_2|$ , we have

$$(**) \leq C_2 B_2(f) \frac{1}{\epsilon_w^d} \left(1 + \frac{1}{\epsilon_w}\right)^2 \epsilon_b. \quad (49)$$

Combining the bound for (\*) and (\*\*) and setting  $C := \max(C_1, C_2)$ ,

$$\sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x}) - R_\sigma^* R f(\mathbf{x})| \leq C \max(M_1(f), B_2(f)) \left\{ \epsilon_w + \frac{1}{\epsilon_w^d} \left(1 + \frac{1}{\epsilon_w}\right)^2 \epsilon_b \right\}$$

where  $C$  only depends on  $\mathcal{X}$ ,  $\phi$ ,  $\psi$ ,  $p_w$ ,  $p_b$  but not on  $f$ ,  $\epsilon_w$ ,  $\epsilon_b$ . Setting  $\sigma_w = \epsilon_w^{-1}$  and  $\sigma_b = \epsilon_b^{-1}$  completes the proof.  $\blacksquare$

**Theorem 6** (Analogue of Theorem 3 for Unbounded  $\phi$ ). *In the same setting of Theorem 3, but with the assumption  $\phi \in C_0^*(\mathbb{R})$  replaced with  $\phi \in C_1^*(\mathbb{R})$ , for any  $f \in C^1(\mathbb{R}^d)$  with  $M_1^*(f) < \infty$ , with probability at least  $1 - \delta$ ,*

$$\sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x}) - I_{\sigma, D, N} f(\mathbf{x})| \leqslant CM_1^*(f) \left\{ \frac{1}{\sigma_{\mathbf{w}}} + \frac{\sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)^2}{\sigma_b} + \frac{\sigma_b (\sigma_b + 1) \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)^2}{D^{\frac{1}{d}}} + \frac{\sigma_b \sigma_{\mathbf{w}}^{d+1}}{\delta \sqrt{N}} \right\}$$

where  $C$  is a constant that may depend on  $\mathcal{X}, \mathbb{1}, \phi, \psi, p_{\mathbf{w}}, p_b$  but does not depend on  $f, \sigma_{\mathbf{w}}, \sigma_b, \delta$ .

**Proof** This result follows from a modification of the proof of Theorem 3. Recall from (36) there are three terms, (\*), (\*\*), and (\*\*\*) , to be bounded. In what follows we indicate how the arguments used to establish Theorem 3 should be modified.

**Bounding (\*)**: Replace (38) in the proof of Theorem 3 with the following inequality

$$\begin{aligned} B_2(f_{\text{clip}}) &= \int_{[-S, S]^d} f_{\text{clip}}(\mathbf{x}) (1 + \|\mathbf{x}\|_2)^2 d\mathbf{x} \\ &\leqslant \int_{[-S, S]^d} M_1(f_{\text{clip}}) \cdot (1 + S)^2 d\mathbf{x} = (1 + S)^2 (2S)^d M_1(f_{\text{clip}}) < \infty. \end{aligned}$$

Now apply Theorem 5 in place of Theorem 2 to obtain

$$(*) \leqslant C'_1 \max(M_1(f_{\text{clip}}), B_2(f_{\text{clip}})) \left( \frac{1}{\sigma_{\mathbf{w}}} + \frac{\sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)^2}{\sigma_b} \right)$$

for some constant  $C'_1 > 0$  depending only on  $\phi, \psi, p_{\mathbf{w}}, p_b$ . Let  $C_1 = (1 + S)^2 (2S)^d C'$  to see

$$(*) \leqslant C_1 M_1(f_{\text{clip}}) \left( \frac{1}{\sigma_{\mathbf{w}}} + \frac{\sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)^2}{\sigma_b} \right).$$

**Bounding (\*\*)**: Here we replace the upper bound of (b) in the proof of Theorem 3. From (45),

$$\begin{aligned} \text{(b)} &= Z \int_{\mathbb{R}} \int_{\mathbb{R}^d} \left| (1 + \|\mathbf{w}\|_2) \phi(\mathbf{w} \cdot \mathbf{x} + b) p_{\mathbf{w}, \sigma}(\mathbf{w}) p_{b, \sigma}(b) \right| d\mathbf{w} db \\ &\leqslant Z C_{\phi} (1 + \|\mathbf{x}\|_2) \underbrace{\int_{\mathbb{R}^d} (1 + \|\mathbf{w}\|_2)^2 p_{\mathbf{w}, \sigma}(\mathbf{w}) d\mathbf{w}}_{\text{(c)}} \underbrace{\int_{\mathbb{R}} (1 + |b|) p_{b, \sigma}(b) db}_{\text{(d)}}. \end{aligned}$$

Let  $C_{p_{\mathbf{w}}} := \max(1, 2 \int_{\mathbb{R}^d} \|\mathbf{w}\|_2 p_{\mathbf{w}}(\mathbf{w}) d\mathbf{w}, \int_{\mathbb{R}^d} \|\mathbf{w}\|_2^2 p_{\mathbf{w}}(\mathbf{w}) d\mathbf{w})$  and  $C_{p_b} := \max(1, \int_{\mathbb{R}} |b| p_b(b) db)$ . By the same discussion on the upper bound of (b) in the proof of Theorem 3,

$$\text{(c)} \leqslant C_{p_{\mathbf{w}}} (1 + \sigma_{\mathbf{w}})^2, \quad \text{(d)} \leqslant C_{p_b} (1 + \sigma_b).$$

Recall  $Z = (2\pi)^{\frac{1}{2}} \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \|\widehat{p_b}\|_{L^1}^{-1} \sigma_{\mathbf{w}}^d \sigma_b$ . By the upper bound of (c) and (d),

$$\text{(b)} \leqslant (1 + \|\mathbf{x}\|_2) (2\pi)^{\frac{1}{2}} \|\widehat{p_{\mathbf{w}}}\|_{L^1}^{-1} \|\widehat{p_b}\|_{L^1}^{-1} C_{p_{\mathbf{w}}} C_{p_b} C_{\phi} \sigma_b (1 + \sigma_b) \sigma_{\mathbf{w}}^d (1 + \sigma_{\mathbf{w}})^2.$$

Let  $C'_2 := \sup_{\mathbf{x} \in [-S, S]^d} |(1 + \|\mathbf{x}\|_2)|(2\pi)^{\frac{1}{2}} \|\widehat{p}_{\mathbf{w}}\|_{L^1}^{-1} \|\widehat{p}_b\|_{L^1}^{-1} C_{p_{\mathbf{w}}} C_{p_b} C_\phi$ . By plugging the upper bound of (b) in (39) and setting  $C_2 := C_Q M_1(\psi) C'_2$ , we have

$$(**) \leq C_2 M_1(f_{\text{clip}}) \frac{\sigma_b(\sigma_b + 1) \sigma_{\mathbf{w}}^d (\sigma_{\mathbf{w}} + 1)^2}{D^{\frac{1}{d}}}.$$

**Bounding (\*\*):** Finally we apply Lemma 5 in place of Lemma 3. From (40),  $\mathbb{E}_{\{\mathbf{w}_i, b_i\}_{i=1}^N} [(**)]$  is upper bounded by

$$\mathbb{E}_{\{\mathbf{w}_i, b_i\}_{i=1}^N} [(**)] \leq \frac{C_{\mathcal{X}}(2S)^d M_1(f_{\text{clip}}) \|\psi\|_{L^\infty} L_\phi V_{p_{\mathbf{w}}} \sigma_{\mathbf{w}}}{\sqrt{N}}$$

Set  $C_3 := C_{\mathcal{X}}(2S)^d \|\psi\|_{L^\infty} L_\phi V_{p_{\mathbf{w}}} \sigma_{\mathbf{w}}$ . By Lemma 5, with probability at least  $1 - \delta$ ,

$$(***) \leq C_3 M_1(f_{\text{clip}}) \frac{\sigma_b \sigma_{\mathbf{w}}^{d+1}}{\delta \sqrt{N}}.$$

For all bounds of (\*), (\*\*), and (\*\*\*), recall that  $M(f_{\text{clip}}) \leq 3M_1^*(\mathbb{1})M_1^*(f)$  from (37). Then combining (\*), (\*\*), and (\*\*\*) and setting  $C := 3M_1^*(\mathbb{1})(C_1 + C_2 + C_3)$  completes the proof.  $\blacksquare$

## A.4 Experiments

This section expands on the experiments that were reported in the main text. Appendix A.4.1 reports the cubature rules that were used. Appendix A.4.2 discusses how to measure the similarity between BNNs and GPs by the maximum mean discrepancy (MMD) as an alternative error criteria to the MRMSE and shows the MMD of the same experiment as Section 6.1. Appendix A.4.3 explores the effect of using alternative settings in the ridgelet prior, compared to those used to produce the figures in the main text.

In some instances of the experiments we report, poor numerical conditioning was encountered. The results that we present employed a crude form of numerical regularisation in order that such issues – which arise from the posterior approximation approach used and are not intrinsic to the ridgelet prior itself – were obviated. Specifically, we employed the Moore–Penrose pseudo-inverse whenever the action of an inverse matrix was required, and we employed Tikhonov regularisation (with the extent of the regularisation manually selected) whenever a matrix square root was required.

### A.4.1 EXPERIMENTAL SETTING

Here, for completeness, we report the cubature rules  $\{(u_i, \mathbf{x}_i)\}_{i=1}^D$  and the bandwidth parameters  $\sigma_{\mathbf{w}}, \sigma_b$  that were used in our experiments. The sensitivity of reported results to these choices is investigated in Appendix A.4.3.

Recall that  $\{(u_i, \mathbf{x}_i)\}_{i=1}^D$  is a cubature rule on  $[-S, S]^d$  and that, in all experiments, we aim for accurate approximation on  $\mathcal{X} = (-5, 5)^d$ . For  $S > 5$  a mollifier  $\mathbb{1}$  is required and in this case (13) we used. The settings that we used for the results in the main text were as follows. As aforementioned in Section 6, the functions  $\phi$  and  $\psi$  were set as Table 1 and the densities  $p_{\mathbf{w}}$  and  $p_b$  were set as standard Gaussians for all of the results in Table 2.



Experiment	Cubature points, weights, total number	Bandwidth of $p_w, p_b$
Section 6.1	$x_j$ : grid point on $[-6, 6]$ , $u_j = 12D^{-1}\mathbb{1}(x_j)$ , $D = 200$	$\sigma_w = 3$ , $\sigma_b = 12$
Section 6.2 CO <sub>2</sub>	$x_j$ : grid point on $[-5, 5]$ , $u_j = 10D^{-1}$ , $D = 200$	$\sigma_w = 3$ , $\sigma_b = 12$
Section 6.2 Airline	$x_j$ : grid point on $[-5, 5]$ , $u_j = 10D^{-1}$ , $D = 200$	$\sigma_w = 3$ , $\sigma_b = 12$
Section 6.2 In-painting	$\mathbf{x}_j$ : grid point on $[-5, 5]^2$ , $u_j = 10^2D^{-1}$ , $D = 30^2$	$\sigma_w = 2$ , $\sigma_b = 12$
Section 6.3 Deep BNN	$x_j$ : grid point on $[-6, 6]$ , $u_j = 12D^{-1}\mathbb{1}(x_j)$ , $D = 50$	$\sigma_w = 3$ , $\sigma_b = 12$

Table 2: Cubature rules and bandwidth parameters used for each experiment reported in the main text.

#### A.4.2 ALTERNATIVE ERROR MEASUREMENT BY MMD

In the main text we adopted the MRMSE to measure the similarity between a BNN and the target GP. However, authors such as Matthews et al. (2018) considered instead a two sample test based on MMD. The purpose of this appendix is to present complementary results to those in the main text, based instead on MMD. The MMD is a distance between two probability measures  $\mathbb{P}$  and  $\mathbb{Q}$ , defined as

$$\text{MMD}^2(\mathbb{P}, \mathbb{Q}) := \sup_{\|h\|_{\mathcal{H}} \leq 1} |\mathbb{E}_{Y \sim \mathbb{P}}[f(Y)] - \mathbb{E}_{Y \sim \mathbb{Q}}[f(Y)]|$$

where  $\mathcal{H}$  is a reproducing kernel Hilbert space uniquely associated with a kernel  $K$ ; see e.g. Muandet et al. (2017). The MMD has the closed form

$$\text{MMD}^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{Y, Y' \sim \mathbb{P}}[K(Y, Y')] - 2\mathbb{E}_{Y \sim \mathbb{P}, Y' \sim \mathbb{Q}}[K(Y, Y')] + \mathbb{E}_{Y, Y' \sim \mathbb{Q}}[K(Y, Y')],$$

which can be approximated using samples from  $\mathbb{P}$  and  $\mathbb{Q}$ . To this end, we consider the BNN and GP’s prior predictives on fixed inputs  $\{\mathbf{x}_i\}_{i=1}^M$  that are  $M$ -dimensional probability distributions  $\mathbb{P}$  and  $\mathbb{Q}$  and employ the squared exponential kernel

$$K(\mathbf{y}, \mathbf{y}') := \exp\left(-\frac{1}{\alpha^2}\|\mathbf{y} - \mathbf{y}'\|_2^2\right), \quad \mathbf{y}, \mathbf{y}' \in \mathbb{R}^M$$

with  $1/\alpha^2 = 0.001$ . For the experiment that we report here, we set  $M = 50$  where  $\{x_i\}_{i=1}^{50}$  are a regular grid over  $(-5, 5)$ , and we use 1,000 samples from  $\mathbb{P}$  and  $\mathbb{Q}$  to approximate  $\text{MMD}^2$ , where  $\mathbb{P}$  and  $\mathbb{Q}$  are the BNN and GP’s prior predictives on  $\{x_i\}_{i=1}^{50}$ .

The same experiment as in Section 6.1 was assessed using MMD, with results presented in Figure 10. It can be observed that MMD decreases when  $N$  increases, likewise the MRMSE in Figure 5. The rate of decrease appears somewhat slower than observed in Matthews et al. (2018), but this is to be expected since we are attempting to approximate an arbitrary GP whereas Matthews et al. (2018) consider approximation of the GP defined as the natural infinite-width limit of the BNN.

#### A.4.3 COMPARISON OF PRIOR PREDICTIVE WITH DIFFERENT SETTINGS

In this final section we investigate the effect of varying the settings of the ridgelet prior. The default settings in Table 1 were taken as a starting point and were then systematically varied. Initially we consider a squared exponential covariance model for the target GP. Specifically, we considered:

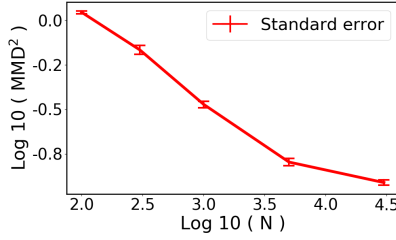


Figure 10: MMD computed from 1000 random samples of GP and BNN outputs at 50 inputs  $\{\mathbf{x}_i\}_{i=1}^{50}$ . The standard error is computed by 10 independent experiments.

1. Different choices of  $\sigma_w$  and  $\sigma_b$ :  $(\sigma_w, \sigma_b) = (1, 2), (2, 6), (3, 12), (4, 20), (5, 30)$
2. Dynamic setting of  $\sigma_w$  and  $\sigma_b$ :  $\sigma_w$  and  $\sigma_b$  varies depending on  $N$
3. Different choices of activation function: Gaussian, Tanh, ReLU
4. Different choices of GP covariance model: squared exponential, rational quadratic, periodic

The findings in each case are summarised next.

**Different choices of  $\sigma_w$  and  $\sigma_b$ :** Figure 11 displays the MRMSE and BNN covariance function for each of the choices  $(\sigma_w, \sigma_b) = (1, 2), (2, 6), (4, 20), (5, 30)$ . Note the choice  $(\sigma_w, \sigma_b) = (3, 12)$  is displayed in Figure 5. It can be observed that the BNN covariance function for larger  $(\sigma_w, \sigma_b)$  has a qualitatively correct shape but is larger overall compared to the GP target when  $N$  is small. On the other hand, the BNN covariance function for smaller  $(\sigma_w, \sigma_b)$  takes values that are closer to that of the GP, but is visually flatter than the GP and the approximation does not improve as  $N$  is increased. These observation indicates that it may be advantageous to change the values of  $(\sigma_w, \sigma_b)$  in a manner that increases with  $N$ . This leads us to the next experiment.

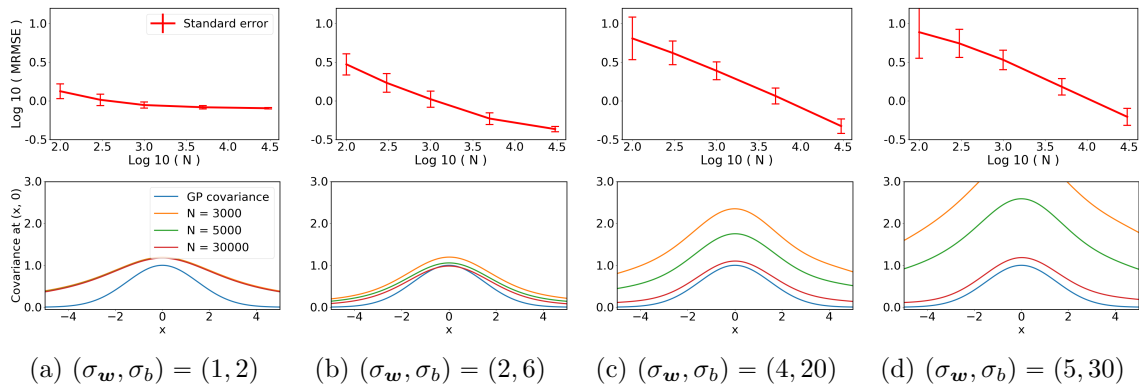


Figure 11: MRMSE and BNN covariance for  $(\sigma_w, \sigma_b) = (1, 2), (2, 6), (4, 20), (5, 30)$ .

**Dynamic setting of  $\sigma_w$  and  $\sigma_b$ :** As observed in the previous experiment, it may be advantageous to change  $(\sigma_w, \sigma_b)$  in a manner that increases with  $N$ . In this experiment, the values of  $(\sigma_w, \sigma_b, N)$  were varied as  $(\sigma_w, \sigma_b, N) = (1, 2, 300), (2, 6, 1000), (3, 12, 3000), (4, 20, 10000), (5, 30, 30000)$ . Figure 12 shows BNN sample paths and Figure 13 shows MRMSE and BNN covariance. This dynamic setting of  $(\sigma_w, \sigma_b, N)$  appears to constitute a promising compromise between the two extremes of behaviour observed in Figure 11.

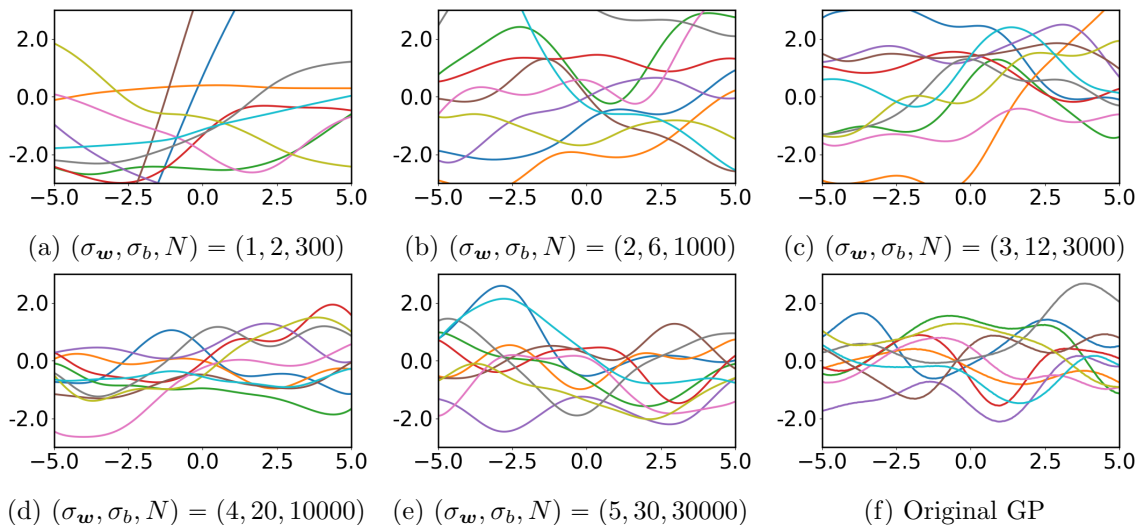


Figure 12: Sample paths of BNN for dynamic setting of  $(\sigma_w, \sigma_b, N)$ .

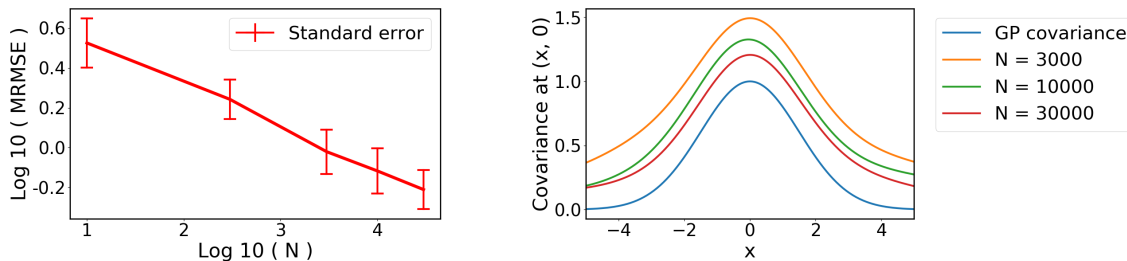


Figure 13: MRMSE and BNN covariance for dynamic setting of  $(\sigma_w, \sigma_b, N)$ .

**Different choices of activation function:** In this experiment, we fix  $(\sigma_w, \sigma_b) = (3, 12)$  and use 3 different activation functions: Gaussian, hyperbolic tangent, and ReLU. The settings for the ridgelet prior corresponding to each activation function are given in Table 1. Figure 14 and Figure 15 indicate that smooth and bounded activation functions, such as the Gaussian activation function, allowed the ridgelet approximation to converge more rapidly to the GP in this experiment.

**Different choice of GP covariance model:** For these experiments we fixed the activation function to the hyperbolic tangent. Then we considered in turn each of the following covariance models for the target GP: square exponential  $k_1$ , rational quadratic  $k_2$ , and

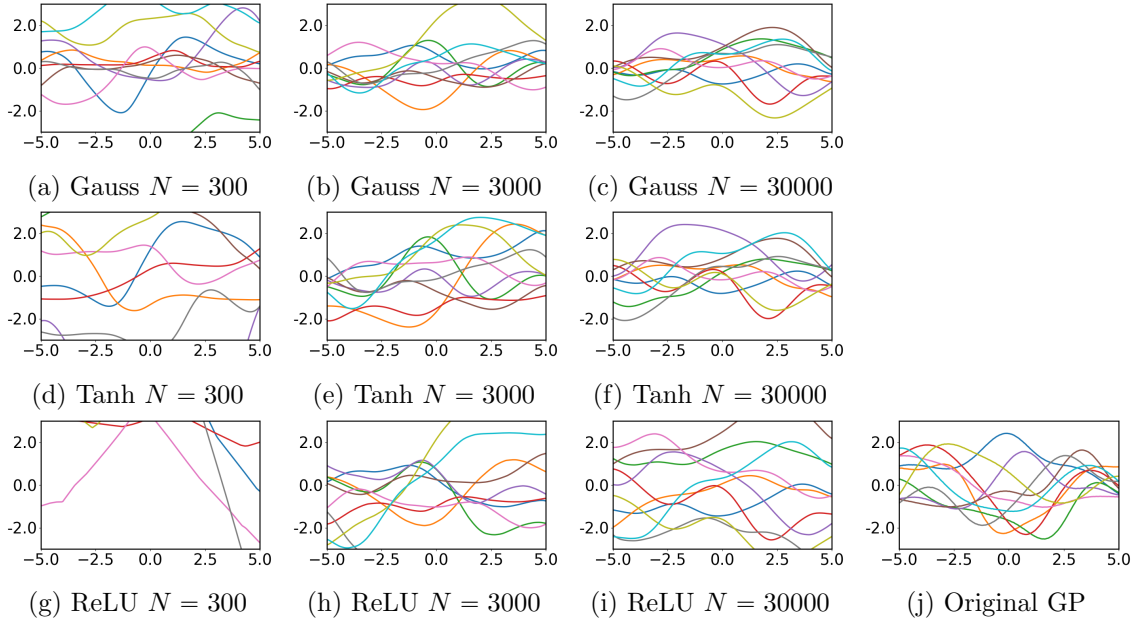


Figure 14: Sample paths of the BNN for different activation functions; Gaussian, hyperbolic tangent, and ReLU.

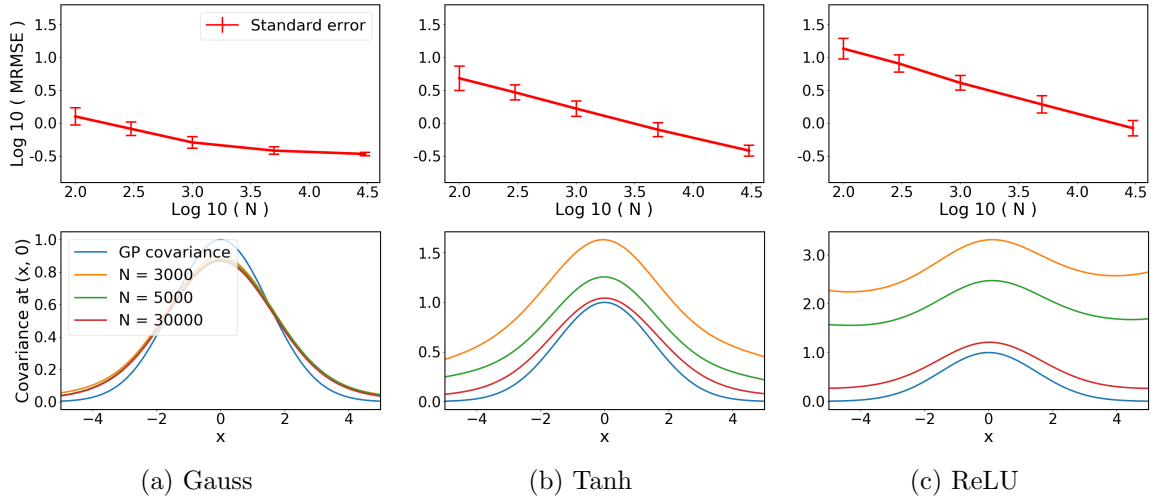


Figure 15: MRMSE and BNN covariance for different activation functions; Gaussian, hyperbolic tangent, and ReLU.

periodic  $k_3$ :

$$\begin{aligned}
 k_1(x, x') &:= l_1^2 \exp\left(-\frac{1}{2s_1^2}|x - x'|^2\right) \\
 k_2(x, x') &:= l_2^2 \left(1 + \frac{1}{2\alpha_2 s_2^2}|x - x'|^2\right)^{-\alpha_2} \\
 k_3(x, x') &:= l_3^2 \exp\left(-\frac{5}{2} \frac{\sin\left(\frac{\pi}{p_3}|x - x'|\right)^2}{s_3^2}\right)
 \end{aligned}$$

where  $l_1 = 1.0, s_1 = 1.5, l_2 = 1.0, \alpha_2 = 1.0, s_2 = 0.75, l_3 = 1.0, p_3 = 2.0, s_3 = 0.75$ . For the choice of  $(\sigma_w, \sigma_b, N)$ , we used the five combination  $(\sigma_w, \sigma_b, N) = (1, 2, 300), (2, 6, 1000), (3, 12, 3000), (4, 20, 10000), (5, 30, 30000)$  examined in the previous experiment. Such dynamic choice of  $(\sigma_w, \sigma_b)$  is useful to achieve a better approximation quality when the covariance model is complex. The sample paths from the BNN with the ridgelet prior are displayed in Figure 16 as a function of  $N$ , and the associated BNN covariance functions are displayed in Figure 17. It is perhaps not surprising that the periodic covariance model, being the most complex, appears to be the most challenging to approximate with a BNN.

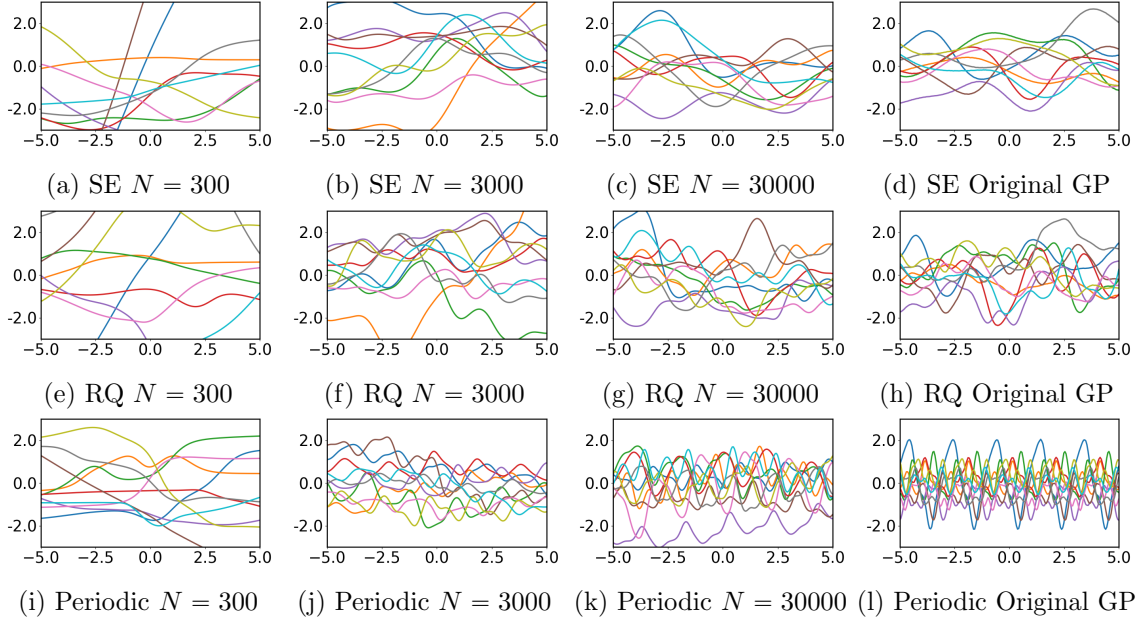


Figure 16: Sample paths of the BNN for different GP covariance models; square exponential, rational quadratic, and periodic.

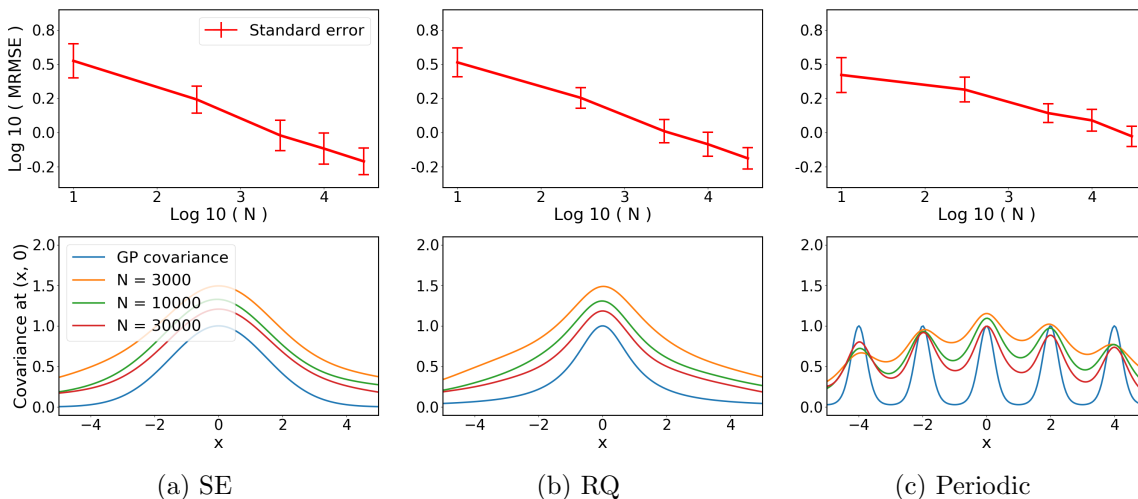


Figure 17: MRMSE and BNN covariance for different GP covariance models; square exponential, rational quadratic, and periodic.

## References

- Ben Adlam, Jaehoon Lee, Lechao Xiao, Jeffrey Pennington, and Jasper Snoek. Exploring the uncertainty properties of neural networks’ implicit priors in the infinite-width limit. *The 9th International Conference on Learning Representations*, 2021.
- Shadi AlZu’bi, Naveed Islam, and Maysam Abbod. Multiresolution analysis using wavelet, ridgelet, and curvelet transforms for medical image segmentation. *International Journal of Biomedical Imaging*, 2011:136034, 2011.
- Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18:1–53, 2018.
- Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.
- Wray L. Buntine and Andreas S. Weigend. Bayesian back-propagation. *Complex Systems*, 5(6):603–643, 1991.
- Emmanuel Candès. *Ridgelets: Theory and Applications*. PhD thesis, Stanford University, 1998.
- George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.
- Andreas C. Damianou and Neil D. Lawrence. Deep Gaussian processes. *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, PMLR 31:207–215, 2013.
- Philip J. Davis and Philip Rabinowitz. *Methods of Numerical Integration*. Academic Press, 1984.
- Minh N. Do and Martin Vetterli. The finite ridgelet transform for image representation. *IEEE Transactions on Image Processing*, 12(1):16–28, 2003.

- Matthew M. Dunlop, Mark A. Girolami, Andrew M. Stuart, and Aretha L. Teckentrup. How deep are deep Gaussian processes? *Journal of Machine Learning Research*, 19(54):1–46, 2018.
- David Duvenaud. *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge, 2014.
- Daniel Flam-Shepherd, James Requeima, and David Duvenaud. Mapping Gaussian process priors to Bayesian neural networks. *The 31st Conference on Neural Information Processing Systems Bayesian Deep Learning Workshop*, 2017.
- Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow Gaussian processes. *The 7th International Conference on Learning Representations*, 2019.
- Evarist Gine and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2015.
- Loukas Grafakos. *Classical Fourier Analysis*. Springer, 2000.
- Danijar Hafner, Dustin Tran, Timothy Lillicrap, Alex Irpan, and James Davidson. Noise contrastive priors for functional uncertainty. *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, PMLR 115:905–914, 2020.
- Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, 2015.
- Lee K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics*, 20(1):608–613, 1992.
- Charles D. Keeling and Timothy P. Whorf. Monthly atmospheric CO<sub>2</sub> records from sites in the sio air sampling network. In *Trends: A Compendium of Data on Global Change*. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, 2004.
- Mohammad Emtiyaz E. Khan, Alexander Immer, Ehsan Abedi, and Maciej Korzepa. Approximate inference turns deep networks into Gaussian processes. *Proceedings of the 33rd Conference on Neural Information Processing Systems*, 2019.
- Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, 2008.
- Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Fixing asymptotic uncertainty of Bayesian neural networks with infinite ReLU features. *arXiv:2010.02709*, 2020.
- Vera Kurková and Marcello Sanguineti. Bounds on rates of variable-basis and neural-network approximation. *IEEE Transactions on Information Theory*, 47(6):2659–2665, 2001.
- Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as Gaussian processes. *The 6th International Conference on Learning Representations*, 2018.
- Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.
- Elijah Liflyand. Integrability spaces for the Fourier transform of a function of bounded variation. *Journal of Mathematical Analysis and Applications*, 436(2):1082–1101, 2016.
- 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.

- David J. C. Mackay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.
- Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *The 6th International Conference on Learning Representations*, 2018.
- John Mitros and Brian Mac Namee. On the validity of Bayesian neural networks for uncertainty estimation. *Proceedings of the 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*, 2019.
- Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.
- Noboru Murata. An integral representation of functions using three-layered networks and their approximation bounds. *Neural Networks*, 9(6):947–956, 1996.
- Iain Murray, Ryan P. Adams, and David J. C. Mackay. Elliptical slice sampling. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 9:541–548, 2010.
- Eric Nalisnick. *On Priors for Bayesian Neural Networks*. PhD thesis, University of California, Irvine, 2018.
- Radford M Neal. *Bayesian Learning for Neural Networks*. Springer, 1995.
- Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are Gaussian processes. *The 7th International Conference on Learning Representations*, 2019.
- Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width ReLU nets: The multivariate case. *The 8th International Conference on Learning Representations*, 2020.
- Tim Pearce, Russell Tsuchida, Mohamed Zaki, Alexandra Brintrup, and Andy Neely. Expressive priors in Bayesian neural networks: Kernel combinations and periodic functions. *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, PMLR 115: 134–144, 2020.
- Arya A. Pourzanjani, Richard M. Jiang, and Linda R. Petzold. Improving the identifiability of neural networks for Bayesian inference. *The 31st Conference on Neural Information Processing Systems Workshop on Bayesian Deep Learning*, 2017.
- Giuseppe Da Prato. *An Introduction to Infinite-Dimensional Analysis*. Springer, 2006.
- Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- Emilio Serrano and Javier Bajo. Deep neural network architectures for social services diagnosis in smart cities. *Future Generation Computer Systems*, 100:122–131, 2019.
- Sho Sonoda and Noboru Murata. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 43(2):233–268, 2017.
- Sho Sonoda, Isao Ishikawa, Masahiro Ikeda, Kei Hagihara, Yoshihiro Sawano, Takuo Matsubara, and Noboru Murata. The global optimum of shallow neural network is attained by ridgelet transform. *The 35th International Conference on Machine Learning Workshop on Theory of Deep Learning*, 2018.



- Michael L. Stein. *Interpolation of Spatial Data Some Theory for Kriging*. Springer, 1999.
- Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.
- Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational Bayesian neural networks. *The 7th International Conference on Learning Representations*, 2019.
- Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.
- Leda Tortora, Gerben Meynen, Johannes Bijlsma, Enrico Tronci, and Stefano Ferracuti. Neuroprediction and ai in forensic psychiatry and criminal justice: A neurolaw perspective. *Frontiers in Psychology*, 11, 2020.
- Loring W. Tu. *An Introduction to Manifolds*. Springer, 2010.
- Ernest O Tuck. On positivity of Fourier transforms. *Bulletin of the Australian Mathematical Society*, 74(1):133–138, 2006.
- Aad van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.
- Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.
- Holger Wendland. *Scattered Data Approximation*. Cambridge University Press, 2005.
- Florian Wenzel, Kevin Roth, Bastiaan S. Veeling, Jakub Światkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119:10248–10259, 2020.
- Christopher K. I. Williams. Computing with infinite networks. *Neural Computation*, 10(1):1203–1216, 1998.
- Greg Yang and Hadi Salman. A fine-grained spectral perspective on neural networks. *The 8th International Conference on Learning Representations*, 2020.
- Wanqian Yang, Lars Lorch, Moritz A Graule, Srivatsan Srinivasan, Anirudh Suresh, Jiayu Yao, Melanie F Pradier, and Finale Doshi-Velez. Output-constrained Bayesian neural networks. *The 36th International Conference on Machine Learning Workshop on Uncertainty and Robustness in Deep Learning*, 2019.
- Jiayu Yao, Weiwei Pan, Soumya Ghosh, and Finale Doshi-Velez. Quality of uncertainty quantification for Bayesian neural network inference. *The 36th International Conference on Machine Learning Workshop on Uncertainty and Robustness in Deep Learning*, 2019.