CONSISTENCY OF EMPIRICAL BAYES AND KERNEL FLOW FOR HIERARCHICAL PARAMETER ESTIMATION

YIFAN CHEN, HOUMAN OWHADI, AND ANDREW M. STUART

Abstract. Hierarchical modeling and learning has proven very powerful in the field of Gaussian process regression and kernel methods, especially for machine learning applications and, increasingly, within the field of inverse problems more generally. The classical approach to learning hierarchical information is through Bayesian formulations of the problem, implying a posterior distribution on the hierarchical parameters or, in the case of empirical Bayes, providing an optimization criterion for them. Recent developments in the machine learning literature have suggested new criteria for hierarchical learning, based on approximation theoretic considerations that can be interpreted as variants of cross-validation, and exploiting approximation consistency in data splitting. The purpose of this paper is to compare the empirical Bayesian and approximation theoretic approaches to hierarchical learning, in terms of large data consistency, variance of estimators, robustness of the estimators to model misspecification, and computational cost. Our analysis is rooted in the setting of Matérn-like Gaussian random field priors, with smoothness, amplitude and inverse lengthscale as hierarchical parameters, in the regression setting. Numerical experiments validate the theory and extend the scope of the paper beyond the Matérn setting.

Contents

Date: May 26, 2020.
2010 *Mathematics Subject Classification.* 65F12 62C10 41A05 35Q62.

1. Introduction

1.1. **Background and Context.** Gaussian process regression (GPR) is important in its own right, and as a prototype for more complex inverse problems in which there is a possibly indirect, nonlinear set of observations. An important reason for the success of GPR in applications is its ability to learn hyperparameters, entering through a hierarchical prior, from data. Learning of these hyperparameters is typically achieved through fully Bayesian (sampling) or empirical Bayesian (optimization) methods. However, new approaches suggested in the machine learning literature use approximation theoretic criteria that can be interpreted as variants of cross-validation. In this paper, we develop an analytical framework for the study of hyperparameter learning, comparing the empirical Bayesian and approximation theoretic approaches.

The setting in which we work is the recovery of function $u^\dagger : D \mapsto \mathbb{R}$ from noiseless evaluations at a set of N points contained in the compact set $D \subset \mathbb{R}^d : y_i = u^\dagger(x_i)$ for $1 \le i \le N$. We introduce the compressed notation $\mathcal{X} := (x_1, \dots, x_N)^\mathsf{T}$ and $y = u^\dagger(\mathcal{X}) := (u^\dagger(x_1), \dots, u^\dagger(x_N))^\mathsf{T}$ and try to find a function which is consistent with this noise-free data; from the perspective of numerical analysis, this formulation can be interpreted as an interpolation problem.

The method which we are interested in using to tackle this problem is GPR; these are also known as kernel methods. Given a family of covariance/kernel functions $K_\theta : D \times D \to \mathbb{R}$ which are parameterized by $\theta \in \Theta$, the prototypical GPR solution then is to approximate $u^\dagger$ with the conditional expectation

$$u(\cdot, \theta, \mathcal{X}) := \mathbb{E}\left[\xi(\cdot, \theta) \mid \xi(\mathcal{X}, \theta) = u^\dagger(\mathcal{X})\right] = K_\theta(\cdot, \mathcal{X})[K_\theta(\mathcal{X}, \mathcal{X})]^{-1} u^\dagger(\mathcal{X}), \tag{1.1}$$

where $\xi(\cdot, \theta) \sim \mathcal{GP}(0, K_\theta)$ is a centered Gaussian process[1](GP) with covariance function K_θ. We extend the compressed notation above so that $K_\theta(\mathcal{X}, \mathcal{X})$ denotes the $N \times N$ dimensional Gram matrix with $(i,j)^{\text{th}}$ entry $K_\theta(x_i, x_j)$. Furthermore we write $K_\theta(\cdot, \mathcal{X})$ for the function mapping D to $\mathbb{R}^N$ with i^{th} component $K_\theta(\cdot, x_i) : D \mapsto \mathbb{R}$. We write explicitly the dependence of the solution on the parameter θ and the observed data positions $\mathcal{X}$.

[1]Recall that the covariance function K_θ of a Gaussian process $\mathcal{GP}(0, K_\theta)$ is the kernel of the integral operator representation of C_θ in the Gaussian prior notation $\mathcal{N}(0, C_\theta)$..

All values of $\theta \in \Theta$ produce a solution $u(\cdot, \theta, \mathcal{X})$ which agrees with $u^\dagger$ on the data set $\mathcal{X}$. Nevertheless, different choices may behave differently when considering out-of-sample error, known as generalization error in the machine learning context; this is the approximation error beyond the sampled set $\mathcal{X}$. It is natural to view θ as a hierarchical parameter to be selected based on the data $u^\dagger(\mathcal{X})$ at hand to provide accurate and robust generalization error. The paper is focused on the question of how to select the hierarchical parameter to achieve this goal.

1.2. **Two Approaches.** We study two approaches to the question posed at the end of the last subsection, both based on selecting θ as the optimizer of a variational problem. The two approaches lead to different variational problems, thereby providing two different answers to the question of how to select θ.

1.2.1. *Empirical Bayes Approach.* The empirical Bayes (EB) answer to the above question is to: (1) formulate a prior distribution on the pair (ξ, θ) by assuming that θ is sampled from a prior distribution and that ξ is then sampled from the conditional distribution of $\xi|\theta$; and (2) find the posterior distribution, namely the distribution on the pair (ξ, θ) conditioned on $\xi(\mathcal{X}) = u^\dagger(\mathcal{X})$, and select the parameter θ maximizing the marginal probability of θ subject to this conditioning. For simplicity, here we work with uninformative priors, which leads to the following objective function

$$\mathsf{L}^{\mathrm{EB}}(\theta, \mathcal{X}, u^\dagger) = u^\dagger(\mathcal{X})^\mathsf{T}[K_\theta(\mathcal{X}, \mathcal{X})]^{-1} u^\dagger(\mathcal{X}) + \log\det K_\theta(\mathcal{X}, \mathcal{X})\,. \tag{1.2}$$

This is twice the negative marginal log likelihood of θ given the data $u^\dagger(\mathcal{X})$. Then, EB chooses the parameter by minimizing this objective/loss function:

$$\theta^{\mathrm{EB}}(\mathcal{X}, u^\dagger) := \operatorname*{argmin}_{\theta\in\Theta} \mathsf{L}^{\mathrm{EB}}(\theta, \mathcal{X}, u^\dagger)\,. \tag{1.3}$$

The EB point estimator can also be understood as the maximum likelihood solution for the parameter θ under the observation data and the postulated statistical model.

1.2.2. *Approximation Theoretic Approach.* Approximation theoretic considerations, however, provide a different answer. Rather than proposing a statistical model, this approach works directly with the approximation error and tries to minimize it. An appropriate cost function d on the space of functions is selected, and the methodology proceeds by selecting θ that minimizes $\mathsf{d}(u^\dagger, u(\cdot, \theta, \mathcal{X}))$. In practice, $u^\dagger$ is not available. However, there are ideas in cross-validation that split $\mathcal{X}$ into training data and validation data, and use the approximation error in validation data to help select the parameter in the learning process with the training data; inspired by this idea, we can turn to optimize objective functions with the form

$$\mathsf{d}(u(\cdot, \theta, \mathcal{X}), u(\cdot, \theta, \pi\mathcal{X}))\,, \tag{1.4}$$

where we write $\pi\mathcal{X}$ for a subset of $\mathcal{X}$ obtained by subsampling a proportion, say one-half, of the elements of $\mathcal{X}$. Thus, π is a projection operator and may be selected according to a coarsening of the grid defining $\mathcal{X}$, in which case the objective function is interpreted as the approximation error across different scales of the function. It is also possible to use a randomized strategy, in which the effective objective function becomes $\mathbb{E}_\pi[\mathsf{d}(u(\cdot, \theta, \mathcal{X}), u(\cdot, \theta, \pi\mathcal{X}))]$. More sophisticated subsampling schemes can be adopted.

In this paper, we focus on a particular choice of d. This choice derives from the Kernel Flow (KF) approach introduced in [20]. For the kernel K_θ, we denote by

$(\mathcal{H}_\theta, \|\cdot\|_{K_\theta})$ the associated *Reproducing Kernel Hilbert Space* (RKHS), such that $\|K_\theta(\cdot, x)\|_{K_\theta}^2 = K_\theta(x,x)$. The objective function in KF is chosen as

$$\mathsf{L}^{\mathrm{KF}}(\theta, \mathcal{X}, \pi\mathcal{X}, u^\dagger) := \frac{\|u(\cdot,\theta,\mathcal{X}) - u(\cdot,\theta,\pi\mathcal{X})\|_{K_\theta}^2}{\|u(\cdot,\theta,\mathcal{X})\|_{K_\theta}^2}\,. \tag{1.5}$$

Intuitively, this measures the discrepancy in the RKHS norm between the GPR solution using the whole data $\mathcal{X}$ and using a subset of the data $\pi\mathcal{X}$; the discrepancy is normalized by the RKHS norm of the former solution. As explained above, we understand the numerator as an estimation of the error $\|u^\dagger - u(\cdot,\theta,\mathcal{X})\|_{K_\theta}^2$. We note that such error estimate, based on comparing solutions obtained on two different grids, is a widely used idea in numerical analysis.

Based on Garlerkin orthogonality (see [20]), the objective function admits the following representation formula that is convenient for numerical computation:

$$\mathsf{L}^{\mathrm{KF}}(\theta, \mathcal{X}, \pi\mathcal{X}, u^\dagger) = 1 - \frac{u^\dagger(\pi\mathcal{X})^T K_\theta(\pi\mathcal{X},\pi\mathcal{X})^{-1} u^\dagger(\pi\mathcal{X})}{u^\dagger(\mathcal{X})^T K_\theta(\mathcal{X},\mathcal{X})^{-1} u^\dagger(\mathcal{X})}\,. \tag{1.6}$$

Then, the KF estimator is defined by

$$\theta^{\mathrm{KF}}(\mathcal{X}, \pi\mathcal{X}, u^\dagger) := \operatorname*{argmin}_{\theta\in\Theta} \mathsf{L}^{\mathrm{KF}}(\theta, \mathcal{X}, \pi\mathcal{X}, u^\dagger)\,. \tag{1.7}$$

The KF loss function for GPR favors θ that produces small discrepancy between methods using the whole data set and those using only a fraction of it. For more discussions we refer to [20].

1.2.3. *Guiding Observations.* We observe that EB and KF have distinct objectives: EB seeks to estimate the most likely parameters of the distribution assumed to generate the pointwise observations, while KF chooses parameters to minimize an estimate of the approximation error in the parameter-dependent RKHS norm, and thereby targets the approximation efficiency of the underlying function. EB is inherently probabilistic; KF need not be. It is natural to ask: Are these two approaches related? Do they lead to meaningful estimates of the parameter as $N \to \infty$ for a well-specified statistical model, for example, assuming that $u^\dagger$ is drawn from a GP $\xi(\cdot, \theta^\dagger)$ for some $\theta^\dagger$? How should the estimators be interpreted when the main objective is the approximation of the function $u^\dagger$, rather than the estimation of $\theta^\dagger$? How does this function approximation behave when the model is misspecified, i.e., the true function is not a sample from one of the GPs $\xi(\cdot,\theta)$? The answer to these questions rely on the specification of models for $u^\dagger$ and K_θ, and also the criteria used to define "good" θ for specific applications. We will address aspects of these questions for certain concrete models in this paper.

1.3. **Our Contribution.** In the first part of this work, we study regression based on a Gaussian process prior of the form

$$\mathcal{N}\Big(0, \sigma^2(-\Delta + \tau^2 I)^{-s}\Big), \tag{1.8}$$

with boundary conditions specified on the Laplacian on bounded domain D in order to fully define the covariance; this is a Matérn-like process in which $\theta = (\sigma, \tau, s)$ quantify the amplitude, inverse lengthscale, and regularity of the process, respectively. We study determination of s, σ and τ for the function $u^\dagger$ underlying the data, working with the GP model (1.8). Under this model draws $u^\dagger$ have almost sure Hölder and Sobolev regularity up to $s - d/2$ in the setting that the domain

D is a torus. When the regression points follow a fixed equidistributed design, and $u^\dagger$ is a draw from the Matérn-like process with an unknown regularity $s > d/2$ and known fixed $\sigma = 1, \tau = 0$, we achieve the following theoretical results::

- For the EB method, we prove that the minimizer of the objective function recovers s, asymptotically as the number of regression points tends to infinity.
- For the KF method, we prove that the minimizer of the objective function recovers $\frac{s-d/2}{2}$, asymptotically as the number of regression points tends to infinity.

Furthermore, we make numerical observations about the recovery of the amplitude parameter σ and inverse lengthscale parameter τ for the EB and KF methods. For learning σ, we also prove the large data consistency of EB.

For the regularity parameter, as we see, the EB and KF methods exhibit different large data consistency, therefore leading to different regression functions, on the same data set. The phenomenon suggests interesting questions about the optimal way to approach hierarchical learning in the context of regression. Then in the second part of the work, guided by this observation, we present numerical methods which exemplify, and extend the range of, the first part. We add discussions to compare the two approaches in terms of consistency, variance, robustness of the estimators in model misspecification, and computational efficiency; a summary of our observations are stated below:

- Consistency: For recovery of the regularity parameter, we perform numerical experiments to explore the implication of selecting $t = s$ or $t = \frac{s-d/2}{2}$ in terms of the L^2 approximation error between $u(\cdot, \theta, \mathcal{X})$ and $u^\dagger$. Both of the two parameter choices have non-trial implications, and EB, KF exhibit these implications in the large data limit.
- Variance: we compare the variance of EB, KF estimators numerically, and find that EB leads to a smaller variance.
- Robustness: we numerically investigate settings beyond the Matérn-like process model for $u^\dagger$ and also for K_θ. We define $u^\dagger$ using GPs with different covariance functions, solutions to SPDEs, and deterministic Green functions. We choose K_θ to be the Green's function of different differential operators, where θ encodes information beyond the amplitude, lengthscale, and regularity of the field; for example we consider a setting where it can be the position of a single discontinuity within a conductivity field defined on a one-dimensional domain. We study the behavior of the two estimators in those settings with or without model misspecification. Here, misspecification refers to the mismatch of the model for $u^\dagger$ and the kernel family K_θ.
- Computation: we discuss and make comments about the computational aspects of the two estimators, including their amenability to application of subsampling methods for dimension reduction.

We note that our work on EB for the Matérn-like process model is significantly different from [14], which also demonstrates the identification of regularity s. Our work does not require simultaneous diagonalization of covariance and the operator $\mathcal{L}\mathcal{L}^*$, where $\mathcal{L}$ is the forward operator in the inverse problem, here being the pointwise evaluation operator.

1.4. **Literature Review.** We review the literature in this subsection. Several fields are of relevance, and we label them to organize the review clearly.

Regression and Inverse Problems. Regression is a form of inverse problem [6], and if formulated in a Bayesian fashion, it falls within the scope of Bayesian nonparametric estimation [10, 11]. In the paper [15] a simple class of linear inverse problems was studied from the perspective of posterior consistency, and it was demonstrated that the rate of posterior convergence depends sensitively on the relationship between regularity of the true function being sought, and the regularity of draws from the prior. This motivates the need for hierarchical procedures that adapt, on the basis of the data, the regularity of draws from the prior. In [14] the work in [15] was extended to cover the data-adapted learning of the regularity parameter in the prior; as the authors note: theoretical work "that supports the preference for empirical or hierarchical Bayes methods does not exist at the present time, however. It has until now been unknown whether these approaches can indeed robustify a procedure against prior mismatch. In this paper, we answer this question in the affirmative." This analysis, however, requires simultaneous diagonalization of a self-adjoint operator formed from the forward model and the covariance operator, for all values of the hyper-parameter. Consistency is studied without this assumption in [33], and extended to the study of emulation within Bayesian inversion in [26] and to empirical Bayesian procedures in [27]. The papers [14] and [27] also use the EB loss function (1.2).

Kernel Flow and Cross-validation. The KF loss function in (1.6) was originally derived in [20]. It can be interpreted, from a numerical homogenization perspective [19], as the relative energy contained in the fine scales (in the unresolved part) of $u^\dagger$. In the paper [20], the proposed loss function to be optimized (via SGD) is actually

$$\mathbb{E}_{\pi_1}\mathbb{E}_{\pi_2}\mathsf{L}^{\mathrm{KF}}(\theta, \pi_1\mathcal{X}, \pi_2\pi_1\mathcal{X}, u^\dagger)\,, \tag{1.9}$$

where $\pi_1\mathcal{X}$ is a subsampling of $\mathcal{X}$, and $\pi_2\pi_1\mathcal{X}$ is a further subsampling of $\pi_1\mathcal{X}$. This choice reduces the dimension of the kernel matrix and enables fast computation per iteration. Although the KF loss appears to be new, it can be seen as a variant of cross-validation (CV), which is a commonly used model selection/parameter estimation criteria [1, 9, 16]. A theoretical understanding of the consistency of CV "is very much of interest" [35] since its convergence rate can be shown to be asymptotically minimax [25] or near minimax optimal [28, 30] while having a lower computational complexity [38] than MLE (maximum likelihood estimation). The consistency of parameter estimation for the Ornstein-Uhlenbeck process has been studied in [36] for MLE, and [4] for CV.

In the setting of hyperparameters estimation of GPs, comparing MLE with CV can be traced back to Wahba [31] and Stein [24] who compared variants of these procedures[2] for choosing the smoothing parameter of a smoothing spline; they observed that while MLE is optimal when the model is well-specified, CV may perform better (than MLE) under misspecification (see also [3] for theoretical analysis and [32] for a practical example involving real data [32]) and has a comparable rate of convergence when the model is correct (Stein [24] observed that "both estimates are asymptotically normal with the CV estimate having twice the asymptotic variance

[2]modified maximum likelihood estimation and generalized cross validation

of the MLE estimate" and suggested that "The penalty for using CV instead of MLE when the stochastic model is correct is greater for higher-order smoothing splines, both in terms of the efficiency in estimating the smoothing parameter and the impact on subsequent predictions"). We also refer to [17] for a detailed numerical comparison between MLE and CV for estimating spline smoothing parameters. As observed in [23], these comparisons "are relevant for both numerical analysts and statisticians" since kernel interpolation can be interpreted as both approximating a deterministic unknown function from quadrature points or as estimating a sample from a Gaussian process from pointwise measurements.

Machine Learning and Kernel Learning. Kernel methods and GP have long been used in machine learning [12, 21]. Learning a good kernel for a given task is very important in practice. Many works have tried to learn a kernel from data based on different criteria; for example, in [2], the kernel is modified to make the model have a large margin in classification, and in [5], the kernel is selected to have a small local Rademacher complexity. EB and KF loss functions in this paper can also serve as the criteria and have been used in [21, 34, 20].

The recent discovery of the neural tangent kernel regime for hyperparameterized models [13] also suggests that a theoretical understanding of kernel selection may lead to improved parameter selection for such models. To describe this connection consider a model with input x, parameters θ and output $f(x;\theta)$. The popular machine learning approach to learning θ from the input/output data $(\mathcal{X},\mathcal{Y})$ (with $\mathcal{X}=(x_1,\dots,x_N)^T$ and $\mathcal{Y}=(y_1,\dots,y_N)^T$) is to introduce a loss defined as an empirical risk, i.e.

$$\mathsf{L}(\theta)=\frac{1}{N}\sum_{i=1}^{N} l(f(x_i;\theta),y_i) \tag{1.10}$$

where l is typically a squared error $l(y',y)=|y'-y|^2$ (or a cross-entropy term in classification problems). In the over-parameterized regime (when the dimension of θ is much larger than that of $\mathcal{Y}$, e.g. for wide neural networks of any depth), under mild assumptions [13, 18], learning θ by applying gradient descent or stochastic (sub-sampled) gradient descent to (1.10) is, in a limiting process that can be made precise, equivalent to regressing the data with a specific Gaussian process ξ. If $\theta=(\theta_1,\dots,\theta_I)$ are the parameters of the neural network then $\xi(x)=f(x,\theta)+\sum_i \xi_i\partial_{\theta_i}f(x,\theta)$ (where the ξ_i are i.i.d. $\mathcal{N}(0,1)$ random variables) with covariance function $K_\theta(x,x')=\sum_i \partial_{\theta_i}f(x,\theta)\partial_{\theta_i}f(x',\theta)$ (known as the neural tangent kernel and whose parametric dependence is fixed at the initialization of θ in the gradient descent algorithm). Therefore, from this perspective, it might be fruitful to consider the problem of training such models as that of learning an underlying kernel (by minimizing nonlinear functionals of the empirical distribution such as (1.2) or (1.6)) rather than that of simply fitting the data (by minimizing a generalized moment, i.e., a linear functional, of the empirical distribution such as (1.10)). Numerical experiments presented in [37] appear to suggest that this point of view could improve test errors, generalization gaps, and robustness to distribution shifts. This further motivates the desire to understand the KF-based estimation of θ.

1.5. **Organization.** The organization of the paper is as follows. In Section 2, we discuss the recovery of the regularity parameter in the Matérn-like process.

Subsections 2.1, 2.2 describe the set-up we will base our analysis on, and presents the main results and observations in this paper. To facilitate proof of the main theorem, we provide the Fourier series characterization of $u(\cdot, t, q)$ in Subsection 2.3, which serves as the basis for subsequent discussions on establishing tight estimates of the loss functions. The main representation result in this section is Theorem 2.5. Relying on the representation, we prove the consistency of the EB estimator in Subsection 2.4, and the KF estimator in Subsection 2.5. The motivation and the intuitive ideas behind the flow of the proof are explained in the corresponding sections; detailed proofs are collected at the end of the paper. In Section 3, we provide a numerical demonstration to support and extend the content of this paper. Subsection 3.1 studies whether the two methods can recover the amplitude and lengthscale parameter of the Matérn-like process. Subsection 3.2 discusses the variance of the estimator. Subsection 3.3 discusses well-specified models beyond the Matérn-like process, while Subsection 3.4 is devoted to model misspecifications. Subsection 3.5 considers the computational efficiency of the estimators. Section 4 summarizes the paper and discusses potential future work. The final Section 5 presents all the proofs in this paper.

2. Regularity Parameter Recovery for Matérn-like Process

This section is devoted to the theoretical and numerical study of recovering the regularity parameter s of a Matérn-like process. The large data limits of EB and KF estimators are identified and their implications are discussed. In Subsection 2.1 we introduce the precise setting in which the main theorem and numerical results of this section are developed. In Subsection 2.2 we state, discuss and provide numerical experiments illustrating, our main theorem. Subsection 2.3 contains some underpinning results from Fourier analysis that are used in our proofs. Subsections 2.4 and 2.5 contain roadmaps for the proofs of the two constituent parts of the main theorem, with pointers to the appendix for full details.

2.1. **Set-up.**

The general set-up is the same as in Subsection 1.1 in the introduction. In the following subsections we specify the physical domain D for the Laplacian used to define the Gaussian process modeling the function $u^\dagger$, the parameter θ, the kernel function K_θ, the data location $\mathcal{X}$, the loss functions and the estimators derived from them.

2.1.1. *The Physical Domain.* The domain D under consideration is $\mathbb{T}^d = [0,1]^d_{\text{per}}$, the d dimensional unit torus. The space of square integrable functions on $\mathbb{T}^d$ with mean 0 is denoted by

$$\dot{L}^2(\mathbb{T}^d) := \left\{ v : \mathbb{T}^d \to \mathbb{R} \mid \int_{\mathbb{T}^d} |v(x)|^2 \, \mathrm{d}x < \infty, \int_{\mathbb{T}^d} v(x) \, \mathrm{d}x = 0 \right\}. \tag{2.1}$$

The standard L^2 inner product and norm are denoted by $[\cdot, \cdot]$ and $\|\cdot\|_0$ respectively.

2.1.2. *Function Spaces and Stochastic Process Regularity.* To describe the model for $u^\dagger$ and the kernel, we define some useful function spaces; the regularity of Gaussian random fields will be discussed in terms of membership of these spaces.

Let $-\Delta$ be the negative Laplacian equipped with periodic boundary conditions on $\mathbb{T}^d$, and restricted to functions which integrate to zero over $\mathbb{T}^d$. The operator has orthonormal eigenfunctions $\phi_m(x) = e^{2\pi i \langle m, x \rangle}$ with the corresponding eigenvalues $\lambda_m = 4\pi^2 |m|^2$ for $m \in \mathbb{Z}^d \backslash \{0\}$, where $\mathbb{Z}^d$ denotes the d-fold tensor product of the

set of non-negative integers $\mathbb{Z}$. Based on the spectral information of $-\Delta$, we can write each function in $\dot{L}^2(\mathbb{T}^d)$ as a Fourier series

$$v(x) = \sum_{m\in\mathbb{Z}^d} \hat{v}(m) e^{2\pi i\langle m,x\rangle}\,, \tag{2.2}$$

where $\hat{v} : \mathbb{Z}^d \to \mathbb{R}$ satisfies $\hat{v}(0) = 0$ and $\hat{v}(m) = [v, \phi_m]$ for $m \in \mathbb{Z}^d\backslash\{0\}$. For every $t > 0$, we can define a Sobolev-like space $\dot{H}^t(\mathbb{T}^d) \subset \dot{L}^2(\mathbb{T}^d)$, which consists of functions with bounded $\|\cdot\|_t$ norm defined as follows:

$$\|v\|_t^2 := \sum_{m\in\mathbb{Z}^d} (4\pi^2|m|^2)^t |\hat{v}(m)|^2 < \infty\,. \tag{2.3}$$

We set $\dot{H}^0(\mathbb{T}^d) = \dot{L}^2(\mathbb{T}^d)$ and for $t < 0$, the space $\dot{H}^t(\mathbb{T}^d)$ is defined through duality. The Hilbert scale of function spaces defined through varying t serves as the basic ingredient to model the regularity of a function. Note that $-\Delta$ has domain $\dot{H}^2(\mathbb{T}^d)$ and range $\dot{L}^2(\mathbb{T}^d)$ and is boundedly invertible as an operator from $\dot{L}^2(\mathbb{T}^d)$ into itself. Indeed we may define $(-\Delta + \tau^2 I)^{-s}$ as a compact operator from $\dot{L}^2(\mathbb{T}^d)$ into itself for any $s > 0$ and $\tau \geq 0$.

With the function spaces and operators defined above, we can then define the Matérn-like process

$$\mathcal{N}\Big(0, \sigma^2(-\Delta + \tau^2 I)^{-s}\Big)\,, \tag{2.4}$$

where s is the regularity, σ is the amplitude, and τ is the inverse lengthscale. Here, to focus on the regularity parameter only, we fix $\sigma = 1$ and $\tau = 0$; fixing them at other positive values would make no difference to the results. The corresponding GP is $\xi \sim \mathcal{N}(0, (-\Delta)^{-s})$. We require $s > d/2$, which ensures the continuity of the sample path of ξ almost surely and guarantees that $\dot{H}^s(\mathbb{T})$ is a RKHS. Our theoretical results regarding the consistency of EB and KF estimators will be based on the assumption that $u^\dagger$ is a draw from the GP $\mathcal{N}(0, (-\Delta)^{-s})$.

We make some notes on the regularity of the GP. The CameronMartin space for $\xi \sim \mathcal{N}(0, (-\Delta)^{-s})$ is $\dot{H}^s(\mathbb{T}^d)$. However, ξ is not an element of the Cameron-Martin space $\dot{H}^s(\mathbb{T}^d)$, almost surely; indeed, we have the property that ξ belongs to $\dot{H}^{s-d/2-\eta}(\mathbb{T}^d)$ for any $\eta > 0$ almost surely (and to Hölder spaces with the same number of fractional derivatives); furthermore, the regularity of the path is spatially homogeneous, i.e., there will be no subdomain $X \subset \mathbb{T}^d$ such that ξ behaves differently in regularity. Here, we refer, for this phenomenon, to ξ having *homogeneous critical regularity* $s - d/2$ across $\mathbb{T}^d$. If we drop the term "homogeneous", we mean the property holds without the requirement of spatial homogeneity.

2.1.3. *The Data.* We collect observational data of a function $u^\dagger : \mathbb{T}^d \mapsto \mathbb{R}$ on the physical domain. We observe equidistributed point-wise values of $u^\dagger$ over the torus. To describe the data locations we introduce a level parameter $q \in \mathbb{N}$, such that, for a given q, we have the data locations $x_j = (j_1, j_2, ..., j_d)\cdot 2^{-q}$ for every $j \in J_q$, where the index set is defined by $J_q = \{(j_1, j_2, ..., j_d) \in \mathbb{N}^d : 0 \leq j_k \leq 2^q - 1,\ \forall\ 1 \leq k \leq d\}$. We may also use the simplified notation $x_j = j2^{-q}$ throughout the paper. The data we are given then comprise the values $\{u^\dagger(x_j)\}_{j\in J_q}$, i.e., $\mathcal{X}(q) = \{x_j : j \in J_q\}$.

2.1.4. *The Estimators.* The kernel function for this problem is chosen as

$$K_\theta(x,y) = [\delta(\cdot - x), (-\Delta)^{-t}\delta(\cdot - y)]\,,$$

where $\delta(\cdot - x)$ is the Dirac function centered at x. Equivalently speaking, K_θ is the Green function of the differential operator $(-\Delta)^t$, with the parameter $\theta = \{t\}$. We introduce a number $\delta > 0$ such that the domain of the parameter $t \in [d/2 + \delta, 1/\delta]$; δ is an arbitrary positive number, and we introduce this compactification of the parameter domain in order to simplify the analysis. The reader should not confuse real number δ with Dirac delta function δ.

Following Subsection 1.2, we can define the EB and KF objective functions, and furthermore the EB and KF estimators. Here, we adapt several notations from Subsection 1.2 to this specific context, by writing t instead of θ, and q instead of $\mathcal{X}(q)$, and $K(t,q)$ instead of $K_\theta(\mathcal{X},\mathcal{X})$. These simplified notations make the analysis cleaner to present. Under such convention, the EB and KF estimators for the regularity parameter are as follows:

(2.5)

$$s^{\mathrm{EB}}(q,u^\dagger) = \operatorname*{argmin}_{t\in[d/2+\delta,1/\delta]} \mathsf{L}^{\mathrm{EB}}(t,q,u^\dagger), \quad \mathsf{L}^{\mathrm{EB}}(t,q,u^\dagger) := \|u(\cdot,t,q)\|_t^2 + \log\det K(t,q)\,,$$

(2.6)

$$s^{\mathrm{KF}}(q,u^\dagger) = \operatorname*{argmin}_{t\in[d/2+\delta,1/\delta]} \mathsf{L}^{\mathrm{KF}}(t,q,u^\dagger), \quad \mathsf{L}^{\mathrm{KF}}(t,q,u^\dagger) := \frac{\|u(\cdot,t,q) - u(\cdot,t,q-1)\|_t^2}{\|u(\cdot,t,q)\|_t^2}\,.$$

Above, $u(\cdot,t,q)$ is the mean of the Gaussian process found by conditioning a prior measure $\mathcal{N}(0,(-\Delta)^{-t})$ on observations of $u^\dagger$ at the observation data with level q. The term $\|u(\cdot,t,q)\|_t^2$ equals the term $u^\dagger(\mathcal{X}(q))^{\mathsf{T}}[K_\theta(\mathcal{X}(q),\mathcal{X}(q))]^{-1}u^\dagger(\mathcal{X}(q))$ in equation (1.2); the former is a representation that is more convenient for theoretical analysis of consistency. The formula (1.2) is useful in numerical computation.

The term $K(t,q)$ is the Gram matrix of the covariance operator $(-\Delta)^{-t}$ with the q-level data, i.e. for $j_1, j_2 \in J_q$, the $(j_1,j_2)^{\mathrm{th}}$ entry of $K(t,q)$ is given by $K_{j_1 j_2}(t,q) = [\delta(\cdot - x_{j_1}), (-\Delta)^{-t}\delta(\cdot - x_{j_2})]$. For the KF loss function, the subsampling operator is fixed to be equidistributed subsampling at half the rate ($q \mapsto q-1$); for this reason we omit the dependence of the subsampling operator π in the notation and simply write $q-1$.

2.2. **Main Theorem, Implications, and Proof Technique.**

2.2.1. *Main Theorem.* Our main result is the following theorem regarding the consistency of the two statistical estimators in the large data limit.

Theorem 2.1. *Fix $\delta > 0$. Suppose $u^\dagger$ is a sample drawn from the Gaussian process $\mathcal{N}(0,(-\Delta)^{-s})$. If $s \in [d/2+\delta, 1/\delta]$ then, for the Empirical Bayesian estimator,*

$$\lim_{q\to\infty} s^{\mathrm{EB}}(q,u^\dagger) = s\,;$$

if $\frac{s-d/2}{2} \in [d/2+\delta, 1/\delta]$ then for the Kernel Flow estimator,

$$\lim_{q\to\infty} s^{\mathrm{KF}}(q,u^\dagger) = \frac{s-d/2}{2}\,.$$

In both cases the convergence is in probability with respect to randomly chosen $u^\dagger$.

Remark 2.2. For economy of notation we will drop explicit reference to the dependence of the loss functions and the estimators on $u^\dagger$ in what follows; we will simply write $\mathsf{L}^{\mathrm{EB}}(t,q)$, $\mathsf{L}^{\mathrm{KF}}(t,q)$, $s^{\mathrm{EB}}(q)$, $s^{\mathrm{KF}}(q)$.

The remainder of this subsection is devoted to numerical experiments illustrating the theory, discussion of the implications of the theory, and an overview of the proof techniques we adopt.

2.2.2. *Numerical Illustration of Theory.* We present a numerical example to demonstrate the main theorem, and its consequences for regression. Consider the one dimensional case, i.e., $d = 1$. We set the ground truth $s = 2.5$ and so $\frac{s-d/2}{2} = 1$. The domain is discretized with $N = 2^{10}$ equidistributed grid points. For our first set of experiments we fix the resolution level of the data points to be $q = 9$, i.e., we have 2^9 equidistributed observations of the unknown function $u^\dagger$. In what follows the Laplaican is as defined in Subsection 2.1.2. Given a sample of $u^\dagger$ from $\mathcal{N}(0, (-\Delta)^{-s})$, we form the loss function for the EB and the KF estimators. A single realization of these loss functions is shown in Figure 1.

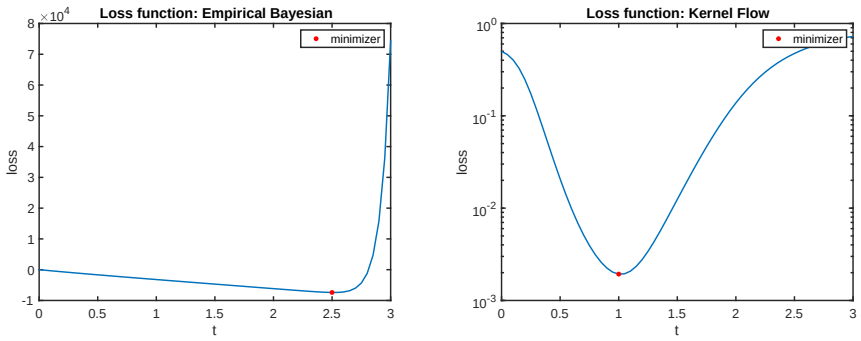

FIGURE 1. Left: EB loss; right: KF loss

We observe that the minimizer of the EB loss function is very close to $t = 2.5$, while the minimizer of the KF loss function is very close to $t = 1$, matching the predictions of Theorem 2.1. Furthermore, the loss functions exhibit some interesting features. Specifically, the EB loss function behaves as a linear function of t, for t less than s, and then blows up rapidly when t exceeds s. The KF loss function is more symmetric with respect to the minimizer $t = \frac{s-d/2}{2}$ in the logarithmic scale. We will make remarks that explain these observations in our theoretical analysis.

We also present here a second set of numerical experiments looking at the effect of the parameter value s selected by EB and KF on the approximation of the function $u^\dagger$, which is (typically) the primary goal of hierarchical parameter estimation. The experimental set-up is the same, but now we vary the resolution of the data points $q = 3, 4, ..., 9$. We study the L^2 error

$$\|u^\dagger(\cdot) - u(\cdot, t, q)\|_0^2 .$$

We start, in Figure 2, by considering the error as a function of q, for different t. As we increase t, the regularity of the Gaussian Process used for regression, or equivalently, the regularity of the basis functions for function approximation, increases. In order to illustrate clear trends, the L^2 error is averaged over the

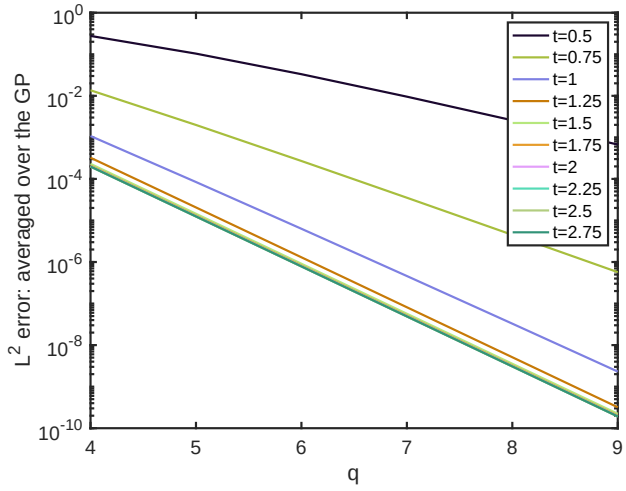

FIGURE 2. L^2 error: averaged over the GP

random draw of $u^\dagger \sim \mathcal{N}(0, (-\Delta)^{-s})$, so the effective error is $\mathbb{E}_{u^\dagger}\|u^\dagger(\cdot) - u(\cdot, t, q)\|_0^2$. From the figure, we can see that when t increases from 0.5 to 1, the convergence rate of the L^2 approximation error increases. Then, if we increase t further from 1 to 3, the slope of the convergence curve remains nearly the same. This demonstrates the fact that $1 = \frac{s-d/2}{2}$ is the minimal t that suffices to achieve the fastest rate of L^2 error convergence. We have observed that this phenomenon is very stable with respect to the specific random draw: the general shape of the curves seen in Figure 2 is still observed when one specific draw of the true random process is used, although the resulting figure contains fluctuations and is not as clear as the average case that we show. Moreover, the observed phenomena also apply for deterministic $u^\dagger$ with homogeneous critical regularity $s - d/2$; for example we observe the same behaviour for

$$u^\dagger(\cdot) = \sum_{m \in \mathbb{Z}^d \setminus \{0\}} \frac{1}{\sqrt{\lambda_m^s}} \phi_m(\cdot)\,,$$

where $\{(\lambda_m, \phi_m)\}_m$ are the eigen-systems of the negative Laplacian, see Subsection 2.1.2.

On the other hand, we can compute $\mathbb{E}_{u^\dagger}\|u^\dagger(\cdot) - u(\cdot, t, q)\|_0^2$ for $q = 9$ as a function of t; see Figure 3. The optimality of the value $s = 2.5$ is clear. However, unlike the experiments in Figure 2, this result is not stable with respect to the random instance of the GP: the minimizer of the L^2 error fluctuates wildly in our experiments.

In summary, the second set of numerical experiments indicates the following implications for the regression accuracy of the EB and KF approaches to hierarchical parameter estimation. The KF estimator selects the minimal t that suffices to achieve the fastest rate of approximation error in the L^2 norm for a given fixed truth; in contrast, the EB estimator converges to the t that achieves the minimal L^2 error, averaged over the draw $u^\dagger \in \mathcal{N}(0, (-\Delta)^{-s})$. In particular, KF is based on purely approximation theoretic considerations whilst EB is founded on statistical considerations as well.

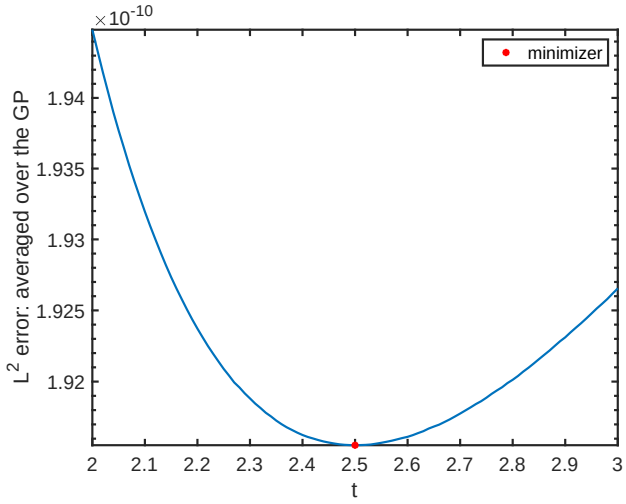

FIGURE 3. L^2 error: averaged over the GP, for $q = 9$

2.2.3. *Further Discussion of The Theory.* We provide some further discussions of the implications of Theorem 2.1 in this subsection. As we see, the theory shows that the EB estimator recovers the ground truth parameter s of the statistical model. This is in line with expectations since the methodology is designed to recover the most likely value of s, given the data, and since the Gaussian measures occurring for different s are mutually singular. In the literature, such consistency results are primarily for observational data in the Fourier domain; thus, the observation operator commutes with the prior. Here, our data model is in the physical domain, which leads to the need for considerably more sophisticated analysis, due to the noncommutativity of the observation operator and the prior operator, and yet is a much more practically useful setting, justifying the investment in the somewhat involved analysis. Our proof provides a novel sharp upper and lower bound on the terms $\|u(\cdot, t, q)\|_t^2$ and $\log\det K(t, q)$, based on techniques in approximation theory and the multiresolution analysis developed in [19]. Our techniques may have broader applications in analyzing the observational model in the physical domain.

Another interesting phenomenon shown in Theorem 2.1 is that the KF estimator, first proposed in [20] as a method to learn kernels for machine learning tasks, achieves a rather different consistency behavior, with the large data limit being $\frac{s-d/2}{2}$. This fact has the following consequence: if the ground truth function $u^\dagger$ has homogeneous critical regularity $s - d/2$, then the KF estimator will converge to half the critical regularity in the large data limit.

To understand the mechanism behind this effect, we observe that the KF loss is a surrogate for the (relative) $\|\cdot\|_t$-norm approximation error between $u^\dagger$ and $u(\cdot, t, q)$. Furthermore, approximation theory implies that the GP regressor $u(\cdot, t, q)$ is also the optimal $\|\cdot\|_t$-norm approximant of $u^\dagger$ in the linear span of the basis functions $\{(-\Delta)^{-t}\delta(x - x_j)\}_{j\in J_q}$. Under this perspective, we see the KF loss incorporates two competing factors in the approximation: increasing t improves the approximation error by increasing the regularity of the basis functions while worsening the measurement of that approximation error by using a stronger norm. The balance

between these two competing factors is achieved when t is half the critical regularity, which is the parameter that KF eventually picks. Our proof provides a detailed demonstration of this phenomenon.

In short, EB learns hierarchically based on statistical principles, whilst KF learns based on approximation theoretic ones. The consistency results presented here provide evidence that the interplay between statistical estimation and numerical approximation can be very useful for parameter estimation and kernel learning in general, thus suggesting new ways of thinking hierarchically. This perspective is one of the main messages that we convey in this paper.

2.2.4. *Proof Strategy.* The following Subsections 2.3, 2.4, 2.5 are devoted to proving the above Theorem 2.1. For the sake of understanding, we provide a high-level view of our proof strategies in this subsection. Fourier analysis plays an important role in the proof. It allows us to analyze the approximation error in a very precise way under this equidistributed design setting.

In our proof, we begin by establishing tight bounds on the terms that appear in the objective functions, i.e., $\|u(\cdot,t,q)\|_t^2$, $\log\det K(t,q)$ and $\|u(\cdot,t,q)-u(\cdot,t,q-1)\|_t^2$, using the toolkit we develop in Subsection 2.3. The norms $\|u(\cdot,t,q)\|_t^2$ and $\|u(\cdot,t,q)-u(\cdot,t,q-1)\|_t^2$ are expressed as random (as a function of $u^\dagger$) series and we carefully analyze the dependencies of the random variables to establish the convergence in probability. For $\log\det K(t,q)$, we employ the multiresolution approach introduced in [19] to establish a tight estimate of the spectrum of the Gram matrix from below and above. Given these estimates, we provide an intuitive understanding of how the loss functions behave and how the minimizers converge in Subsections 2.4, 2.5. In the rigorous treatment, the sharp bounds on the different components of the objective functions will be combined with the uniform convergence result of random series in [29] to obtain the convergence of minimizers.

2.2.5. *Notation.* In many parts of the analysis, we need to develop tight estimates on the terms appearing in the loss functions. Some useful notation for comparing different terms are introduced here. We write $A \simeq B$ if there exists a constant C independent of q,t such that

$$\frac{1}{C}B \le A \le CB\,.$$

The constant may depend on the dimension d and on δ. Correspondingly, if we use $A \gtrsim B$ or $A \lesssim B$, then only one side of the above inequality holds.

Fourier analysis plays a critical role in the analysis. We always use $u^\dagger$ for the ground truth function, while we omit the $\dagger$ symbol for ease of notation when discussing its Fourier transform, and write $\hat{u}$; we will also use $\hat{u}$, with more arguments, to denote the Fourier transform of the Gaussian process mean; see the discussion following Theorem 2.5. In the Fourier domain, we let $B_q := \{m \in \mathbb{Z} : -2^{q-1} \le m \le 2^{q-1}-1\}$ and $B_q^d = B_q \otimes B_q \otimes \cdots \otimes B_q$ be the tensor product of d multiples of B_q. It is a box concentrating around the origin, so only the low-frequency part of the Fourier coefficients are considered.

2.3. **Toolkit: Fourier Series Characterization.** In this subsection, we prepare the necessary tools that are used to prove the main theorem of this paper. We start by establishing a Fourier series characterization for $u(\cdot,t,q)$. This is a key ingredient in expressing the terms in the loss functions as random series. Our approach, using Fourier series, is motivated by the papers [7, 22], where the approximation power

of shift-invariant subspaces of $L^2(\mathbb{R}^d)$ is studied; in our case we use related ideas in the $\dot{L}^2(\mathbb{T}^d)$ setting.

To find the representation of the term $u(\cdot,t,q)$, we invoke its definition, i.e. $u(\cdot,t,q)$ is obtained by GP regression with the q-level data and the covariance function $(-\Delta)^{-t}$. We use the representer theorem from GPR. Concretely, let the set of basis functions be

$$\mathcal{F}_{t,q} = \mathrm{span}_{j\in J_q}\{(-\Delta)^{-t}\delta(\cdot - x_j)\}\,,$$

then, $u(\cdot,t,q)$ is the best approximation in $\mathcal{F}_{t,q}$ to the true function under the $\|\cdot\|_t$ norm. Let us define

$$\hat{\mathcal{F}}_{t,q} := \{g : \mathbb{Z}^d \to \mathbb{C}, \text{there exists an } f \in \mathcal{F}_{t,q} \text{ such that } g = \hat{f}\},$$

the Fourier coefficients of functions in $\mathcal{F}_{t,q}$. A quick observation is that for every $g \in \hat{\mathcal{F}}_{t,q}$, we must have $g(0) = 0$ because of the mean zero property of $f \in \mathcal{F}_{t,q}$. The following proposition gives a complete characterization of the basis functions in $\hat{\mathcal{F}}_{t,q}$, for $t > d/2$.

Proposition 2.3. *For any $g \in \hat{\mathcal{F}}_{t,q}$, there exists a 2^q-periodic function p on $\mathbb{Z}^d$, such that*

$$g(m) = \begin{cases} |m|^{-2t}p(m), & m \neq 0 \\ 0, & m = 0\,. \end{cases}$$

The proof is in Subsection 5.1. Next, we define a 2^q-periodization operator, which will be used to compute the representation of $\hat{u}(m,t,q)$.

Definition 2.4. The operator T_q is defined as a mapping from the space of functions on $\mathbb{Z}^d$ to itself, such that

$$(T_q g)(m) := \sum_{\beta\in\mathbb{Z}^d} g(m + 2^q\beta), \quad m \in \mathbb{Z}^d\,,$$

whenever the right hand side series converges for the function $g : \mathbb{Z}^d \to \mathbb{R}$. We also define

$$M_q^t(m) := \begin{cases} \sum_{\beta\in\mathbb{Z}^d\setminus\{0\}} |2^q\beta|^{-2t}, & \text{if } m = j\cdot 2^q \text{ for some } j \in \mathbb{Z}^d \\ \sum_{\beta\in\mathbb{Z}^d} |m + 2^q\beta|^{-2t}, & \text{else}\,. \end{cases} \tag{2.7}$$

Both $T_q g$ and M_q^t are 2^q-periodic functions on $\mathbb{Z}^d$. Based on this definition, Theorem 2.5 presents the explicit form of the Fourier transform of $u(\cdot,t,q)$; the proof is in Subsection 5.2. The proof relies on the Galerkin orthogonality property of $u(\cdot,t,q)$ due to its being the optimal approximate solution.

Theorem 2.5. *Let $\hat{u}(\cdot,t,q)$ be the Fourier coefficients of $u(\cdot,t,q)$, then for $m \in \mathbb{Z}^d$, we have*

$$\hat{u}(m,t,q) = \begin{cases} 0, & \text{if } m = 0 \\ |m|^{-2t}\frac{(T_q\hat{u})(m)}{M_q^t(m)}, & \text{else} \end{cases}$$

where $\hat{u}$ denotes the Fourier coefficients of $u^\dagger$.

This above representation formula is very useful for analyzing the terms $\|u(\cdot,t,q)\|_t^2$ and $\|u(\cdot,t,q)-u(\cdot,t,q-1)\|_t^2$. As well as studying the Fourier coefficients of $u(\cdot,t,q)$, which we denote by $\hat{u}(\cdot,t,q)$, we will also need to study the Fourier coefficients of $u^\dagger(\cdot)$ which, for ease of notation we will denote by $\hat{u}(\cdot)$, henceforth, omitting the $\dagger$

symbol. It is thus important to look at the number of arguments of $\hat{u}$ to determine which object it is the Fourier transform of. Note also that $u(\cdot,t,q)$ is determined by $u^\dagger$; hence if $u^\dagger$ is random, so is $u(\cdot,t,q)$.

We will use the above Fourier analysis toolkit to study the consistency of EB and KF in the following two subsections.

2.4. **Consistency of the Empirical Bayesian Estimator.** In this subsection, we prove the consistency of the EB estimator. As explained before, our roadmap is to give a tight estimate of the loss functions first and then analyze the minimizers. For the norm term $\|u(\cdot,t,q)\|_t^2$, we invoke Theorem 2.5, based on which this term is expressed as a random series:

Proposition 2.6. *The $\dot{H}^t(\mathbb{T}^d)$ norm of $u(\cdot,t,q)$ has the representation*

$$\|u(\cdot,t,q)\|_t^2 = (4\pi^2)^t \sum_{m\in B_q^d} \frac{|T_q\hat{u}(m)|^2}{M_q^t(m)}\,.$$

Moreover, suppose $u^\dagger \sim \mathcal{N}(0,(-\Delta)^{-s})$ for $s>\frac{d}{2}$, then

$$\|u(\cdot,t,q)\|_t^2 = (4\pi^2)^{t-s} \sum_{m\in B_q^d} \frac{M_q^s(m)}{M_q^t(m)}\xi_m^2\,,$$

where $\{\xi_m\}_{m\in B_q^d}$ are independent unit scalar Gaussian random variables.

Proof. Using Theorem 2.5, we get

$$\begin{aligned}
\|u(\cdot,t,q)\|_t^2 =& \sum_{m\in\mathbb{Z}^d\setminus\{0\}} (4\pi^2)^t |m|^{2t}|\hat{u}(m,t,q)|^2\\
=&(4\pi^2)^t \sum_{m\in\mathbb{Z}^d\setminus\{0\}} |m|^{-2t}\frac{|T_q\hat{u}(m)|^2}{|M_q^t(m)|^2}\\
=&(4\pi^2)^t \sum_{m\in B_q^d} M_q^t(m)\frac{|T_q\hat{u}(m)|^2}{|M_q^t(m)|^2}\\
=&(4\pi^2)^t \sum_{m\in B_q^d} \frac{|T_q\hat{u}(m)|^2}{M_q^t(m)}\,.
\end{aligned}$$

where in the third equality, we use the periodicity of the function $\frac{|T_q\hat{u}(m)|^2}{|M_q^t(m)|^2}$.

If we further assume $u^\dagger\sim\mathcal{N}(0,(-\Delta)^{-s})$, then $\hat{u}(m)\sim\mathcal{N}(0,(4\pi^2)^{-s}|m|^{-2s})$. For different m, these Gaussian random variables are independent. Thus, for different $m\in B_q^d$, we have $T_q\hat{u}(m)\sim\mathcal{N}(0,(4\pi^2)^{-s}M_q^s(m))$, and they are independent. So we can write

$$\sum_{m\in B_q^d} \frac{|T_q\hat{u}(m)|^2}{M_q^t(m)} = (4\pi^2)^{-s}\sum_{m\in B_q^d}\frac{M_q^s(m)}{M_q^t(m)}\xi_m^2\,,$$

where $\{\xi_m\}_{m\in B_q^d}$ are independent unit scalar Gaussian random variables. □

The independence of the random variables established in the preceding representation is crucial for the analysis. The terms $M_q^s(m), M_q^t(m)$ appear in the preceding; to analyze them we present a useful lemma below. The proof is in Subsection 5.3.

Lemma 2.7. *For $t \in [d/2+\delta, 1/\delta]$ and $q \geq 0$, we have*

$$M_q^t(m) \simeq \begin{cases} 2^{-2qt}, & \text{if } m = 0 \\ |m|^{-2t}, & \text{if } m \in B_q^d \backslash \{0\} \end{cases}$$

Moreover, for $m \in B_q^d \backslash \{0\}$, we have $M_q^t(m) - |m|^{-2t} \simeq 2^{-2qt}$.

Now, we are ready to get the estimates of the loss function. The following proposition shows an upper and lower bound on the norm term.

Proposition 2.8 (Bound on the norm term)**.** *Suppose $u^\dagger \sim \mathcal{N}(0, (-\Delta)^{-s})$ for $d/2+\delta \leq s \leq 1/\delta$, then*

$$\|u(\cdot,t,q)\|_t^2 \simeq 2^{-q(2s-2t)}\xi_0^2 + \sum_{m \in B_q^d \backslash \{0\}} |m|^{2t-2s}\xi_m^2 \,,$$

where $\{\xi_m\}_{m \in B_q^d}$ are independent unit scalar Gaussian random variables.

Proof. According to Lemma 2.7, for $m \in B_q^d \backslash \{0\}$, we have $M_q^t(m) \simeq |m|^{-2t}$; for $m = 0$, we have $M_q^t(m) \simeq 2^{-2tq}$. Thus,

$$\begin{aligned} \|u(\cdot,t,q)\|_t^2 &= (4\pi^2)^{t-s} \sum_{m \in B_q^d} \frac{M_q^s(m)}{M_q^t(m)} \xi_m^2 \\ &= (4\pi^2)^{t-s} \left(\sum_{m \in B_q^d \backslash \{0\}} \frac{M_q^s(m)}{M_q^t(m)} \xi_m^2 + \frac{M_q^s(0)}{M_q^t(0)} \xi_0^2 \right) \\ &\simeq 2^{-q(2s-2t)}\xi_0^2 + \sum_{m \in B_q^d \backslash \{0\}} |m|^{2t-2s}\xi_m^2 \,. \end{aligned}$$

This completes the proof. □

Proposition 2.8 states that the behavior of the norm term is nothing but a weighted sum of squares of independent Gaussian random variables, which is amenable to analysis. With this in mind, we state a lemma useful in the analysis of such random series, with proof postponed to Subsection 5.4.

Lemma 2.9. *Suppose $\{\xi_m\}_{m \in \mathbb{Z}^d}$ are independent unit Gaussian random variables.*

- *For $r > 0$, define the random series*

$$\alpha(r,q) = 2^{-qr} \sum_{m \in B_q^d \backslash \{0\}} |m|^{r-d} \xi_m^2 \,.$$

  Fix $\epsilon > 0$, then there exists a function $\gamma(r) > 0$ such that $\lim_{q\to\infty} \alpha(r,q) = \gamma(r) > 0$ uniformly for $r \in [\epsilon, 1/\epsilon]$, where the convergence is in probability.
- *For $r = 0$, define*

$$\alpha(0,q) = \frac{1}{q} \sum_{m \in B_q^d \backslash \{0\}} |m|^{-d} \xi_m^2 \,,$$

  then there exists $\gamma(0) \in (0,\infty)$ such that $\lim_{q\to\infty} \alpha(0,q) = \gamma(0)$ in probability.

We then move to the second term in the loss function, i.e., the log determinant term. It is deterministic and to study it we need a way of analyzing the spectrum of the Gram matrix. The following Proposition 2.10 gives upper and lower bounds on this term. The proof is in Subsection 5.5 and is motivated by analysis developed in the paper [19]. The idea is to use the Schur complement of the Gram matrix and rely on the variational characterization of the Schur complement to get a tight control on the spectrum.

Proposition 2.10 (Bound on the log det term)**.** *For $d/2+\delta \le t \le 1/\delta$, we have*

$$(2t-d)g_1(q) - Cg_2(q) + K(t,0) \le \log\det K(t,q) \le (2t-d)g_1(q) + Cg_2(q) + K(t,0)\,,$$

where $g_1(q) = \sum_{k=1}^{q}(2^{kd} - 2^{(k-1)d})(-k\log 2)$ and $g_2(q) = (2^{qd}-1)(2t-d)$. The constant C is independent of t, q. Moreover, $g_1(q) \simeq -q2^{qd}$.

With the loss function analyzed by the above results, the consistency of the EB estimator is readily stated as follows.

Theorem 2.11 (Consistency of Empirical Bayesian estimator)**.** *Fix $\delta > 0$. Suppose $u^\dagger$ is a sample drawn from the Gaussian process $\mathcal{N}(0, (-\Delta)^{-s})$. If $s \in [d/2+\delta, 1/\delta]$ then*

$$\lim_{q\to\infty} s^{\mathrm{EB}}(q) = s \quad \text{in probability}\,.$$

The detailed proof is in Subsection 5.6. We can understand the theorem intuitively by using the established results above. Recall there are two terms in the loss function: (1) the norm term $\|u(\cdot,t,q)\|_t^2$; (2) the log det term. For the norm term, from Proposition 2.8 and Lemma 2.9, its behavior for $q \to \infty$ is roughly

- Growing like $2^{q(2t-2s+d)}$ if $t > s - d/2$;
- Growing like q if $t = s - d/2$;
- Remaining bounded if $t < s - d/2$.

The log det term decreases like $-(2t-d)q2^{qd}$ according to Proposition 2.10. Noticing that the EB loss function has the form

$$\mathsf{L}^{\mathrm{EB}}(t,q) = \|u(\cdot,t,q)\|_t^2 + \log\det K(t,q)\,,$$

we arrive at the following intuitive observations:

- When $t < s$, the dominant behavior of $\mathsf{L}^{\mathrm{EB}}(t,q)$ is controlled by the log determinant term, since the growth rate of the norm term $2^{q(2t-2s+d)} = o(q2^{qd})$. As a consequence, $\mathsf{L}^{\mathrm{EB}}(t,q)$ exhibits the overall behavior $-(2t-d)q2^{qd}$. Therefore, the loss function decreases linearly with t in this regime. This is consistent with what is observed in Figure 1.
- When $t \ge s$, the increasing speed of the norm term beats the decreasing rate of the log det term, so the norm term dominates the behavior of $\mathsf{L}^{\mathrm{EB}}(t,q)$. Overall, it is like $2^{q(2t-2s+d)}$, which increases exponentially with t; again this is consistent with what is observed in Figure 1.

According to the above observations, the minimizer of $\mathsf{L}^{\mathrm{EB}}(t,q)$ will converge to s. To make the intuition leading to this conclusion rigorous, we need to use techniques of uniform convergence for random series. For details we refer to Subsection 5.6.

2.5. **Consistency of the Kernel Flow Estimator.** In this subsection, we establish the consistency of the KF estimator. As before, we start by estimating the growth behavior of terms that appear in the loss function. We begin with the interaction term $\|u(\cdot,t,q)-u(\cdot,t,q-1)\|_t^2$. Similar to the analysis of the norm term in the preceding subsection, we represent it by using Fourier series.

Proposition 2.12. *The $\dot{H}^t(\mathbb{T}^d)$ norm of $u(\cdot,t,q)-u(\cdot,t,q-1)$ has the representation*

$$\|u(\cdot,t,q)-u(\cdot,t,q-1)\|_t^2=(4\pi^2)^t\sum_{m\in B_q^d}M_q^t(m)\left(\frac{T_q\hat{u}(m)}{M_q^t(m)}-\frac{T_{q-1}\hat{u}(m)}{M_{q-1}^t(m)}\right)^2. \tag{2.8}$$

Proof. By Theorem 2.5, we have

$$\hat{u}(m,t,q)-\hat{u}(m,t,q-1)=\begin{cases}0, & \text{if } m=0\\ |m|^{-2t}\left(\frac{T_q\hat{u}(m)}{M_q^t(m)}-\frac{T_{q-1}\hat{u}(m)}{M_{q-1}^t(m)}\right), & \text{else}.\end{cases}$$

Thus,

$$\begin{aligned}\|u(\cdot,t,q)-u(\cdot,t,q-1)\|_t^2=&(4\pi^2)^t\sum_{m\in\mathbb{Z}^d\setminus\{0\}}|m|^{2t}|\hat{u}(m,t,q)-\hat{u}(m,t,q-1)|^2\\ =&(4\pi^2)^t\sum_{m\in\mathbb{Z}^d\setminus\{0\}}|m|^{-2t}\left(\frac{T_q\hat{u}(m)}{M_q^t(m)}-\frac{T_{q-1}\hat{u}(m)}{M_{q-1}^t(m)}\right)^2\\ =&(4\pi^2)^t\sum_{m\in B_q^d}M_q^t(m)\left(\frac{T_q\hat{u}(m)}{M_q^t(m)}-\frac{T_{q-1}\hat{u}(m)}{M_{q-1}^t(m)}\right)^2.\end{aligned}$$

□

By carefully studying the correlation between the random variables appearing in the preceding proposition, we obtain lower and upper bounds in the following two propositions; proofs can be found in Subsections 5.7 and 5.8.

Proposition 2.13 (Lower bound on the interaction term)**.** *Suppose $u^\dagger\sim\mathcal{N}(0,(-\Delta)^{-s})$ for $d/2+\delta\le s\le 1/\delta$, then*

$$\|u(\cdot,t,q)-u(\cdot,t,q-1)\|_t^2\gtrsim\sum_{m\in B_{q-1}^d\setminus\{0\}}2^{-2tq}|m|^{4t-2s}\xi_m^2,$$

where $\{\xi_m\}_{m\in B_{q-1}^d\setminus\{0\}}$ are independent unit scalar Gaussian random variables.

The upper bound has a more complex form. We introduce the notation $\mathbb{Z}_2^d=\{0,1\}^d$ comprising d dimensional vectors with each component being in $\{0,1\}$. In the following proposition, we also use the convention that $|m|^\alpha=0$ for $m=0$ and any $\alpha\in\mathbb{R}$ to make the notation more compact.

Proposition 2.14 (Upper bound on the interaction term)**.** *Suppose $u^\dagger\sim\mathcal{N}(0,(-\Delta)^{-s})$ for $d/2+\delta\le s\le 1/\delta$, then*

$$\|u(\cdot,t,q)-u(\cdot,t,q-1)\|_t^2\lesssim\sum_{k\in\mathbb{Z}_2^d}\sum_{m\in B_{q-1}^d}(2^{-q(2s-2t)}+2^{-2tq}|m|^{4t-2s})\xi_{k,m}^2,$$

where for a fixed $k\in\mathbb{Z}_2^d$, $\{\xi_{k,m}\}_{m\in B_{q-1}^d}$ are independent unit scalar Gaussian random variables.

We remark that in the upper bound, the random variables for different k may exhibit correlation. However, since the term $\sum_{m\in B^d_{q-1}}(2^{-q(2s-2t)}+2^{-2tq}|m|^{4t-2s})\xi^2_{k,m}$ has the same form for each k, and the number of different k is finite, it suffices to analyze the random series for a single k, in which we have the independence of random variables. The theorem is stated below.

Theorem 2.15 (Consistency of the Kernel Flow estimator)**.** *Fix $\delta > 0$. Suppose $u^\dagger$ is a sample drawn from the Gaussian process $\mathcal{N}(0,(-\Delta)^{-s})$. If $\frac{s-d/2}{2} \in [d/2+\delta, 1/\delta]$ then for the Kernel Flow estimator,*

$$\lim_{q\to\infty} s^{\mathrm{KF}}(q) = \frac{s-d/2}{2} \quad \text{in probability}\,.$$

The idea behind the proof of the theorem is to combine Propositions 2.13, 2.14 and Lemma 2.9. Together they imply the growth behavior of the loss function

$$\mathsf{L}^{\mathrm{KF}}(t,q) = \frac{\|u(\cdot,t,q)-u(\cdot,t,q-1)\|_t^2}{\|u(\cdot,t,q)\|_t^2}$$

as follows:

- When $t < \frac{s-d/2}{2}$, the numerator decays like 2^{-2tq} since $4t-2s<-d$, in which case the summation $\sum_{m\in B^d_q\setminus\{0\}}|m|^{4t-2s}\xi^2_m$ remains bounded. The denominator remains bounded. So the overall behavior is 2^{-2tq}.
- When $\frac{s-d/2}{2} < t < s-d/2$, the numerator decays like $2^{-2tq}\times 2^{q(4t-2s+d)} = 2^{q(2t-2s+d)}$ according to Lemma 2.9. The denominator remains bounded, The overall behavior is $2^{q(2t-2s+d)}$.
- When $t > s-d/2$, the numerator behaves like $2^{q(2t-2s+d)}$, while the denominator behaves like $2^{q(2t-2s+d)}$. The overall behavior is of order 1.

These observations are consistent with what is observed in Figure 1. Based on them we deduce that the minimizer converges to $\frac{s-d/2}{2}$. The loss function exhibits symmetric behavior with respect to $\frac{s-d/2}{2}$ for $t\in(d/2, s-d)$. The detailed rigorous treatment is presented in Subsection 5.9.

3. Numerical Experiments

The setting in Section 2 concerns parameter estimation of functions assumed to be drawn from the measure (1.8). Theorem 2.1, and numerical results reported in Subsection 2.2.2 are concerned entirely with the setting where $\sigma = 1, \tau = 0$ and only s is learned hierarchically, using either the EB or KF approach. This section extends to a wider range of settings, by means of numerical experiments.

In Subsection 3.1 we study the recovery of inverse lengthscale (τ) and amplitude (σ) hyperparameters, not covered by the Theorem 2.1, and not studied in the numerical experiments of Subsection 2.2.2. In Subsection 3.2 we discuss variance of the estimators. In Subsection 3.3, we consider other well-specified models, extending beyond the Matérn-like process example. Subsection 3.4 is devoted to study of both stochastic and deterministic misspecification. In subsection 3.5, we discuss the computational aspects of the EB and KF approaches.

3.1. Recovery of Amplitude and Lengthscale.

An important general principle in looking at the recovery of hyperparameters via EB is to determine whether or not the family of measures are mutually singular with respect to changes in the

parameter to be estimated; learning parameters which give rise to mutually singular families is usually easy, since different almost sure properties can often be used to distinguish measures and this can be achieved without an abundance of data; in contrast those parameters that do not give rise to mutually singular measures typically require an abundance of realizations to be accurately learned. We illustrate this issue in the context of estimating one parameter by EB, the changing of which leads to mutually singular measures, and estimating two parameters by EB, changing one of which leads to mutual singularity, and the other to equivalence, for the Matérn process. We also study analogous questions about identifiability for the KF method. In all cases we work with loss functions that are natural generalizations of (2.5), (2.6).

3.1.1. *Recovery of σ.* A first observation is that the KF loss function is invariant under change of σ, so it cannot recover this parameter. We also note that measures are mutually singular with respect to changes in σ, and so we do expect to be able to recover σ by EB. For the EB estimator, we design the experiment as follows. We study whether the EB method can recover σ while s, τ are fixed. In detail, we consider a problem with domain the one dimensional torus $\mathbb{T}^1$. The Matérn-like kernel has regularity $s = 2.5$, amplitude $\sigma = 1$ and lengthscale $\tau = 0$. We assume the values of s, τ are known, but not σ. We want to recover σ by seeing a single discretized realization $u^\dagger \sim \mathcal{N}(0, \sigma^2(-\Delta + \tau^2 I)^{-s})$. The domain $\mathbb{T}^1$ is discretized into $N = 2^{10}$ equidistributed grid points. The data we observe is the values of $u^\dagger$ in 2^9 equidistributed points. We build the EB loss function (see equation (3.1)) and plot the figure for a single instance; see Figure 4. We introduce ς as the variable

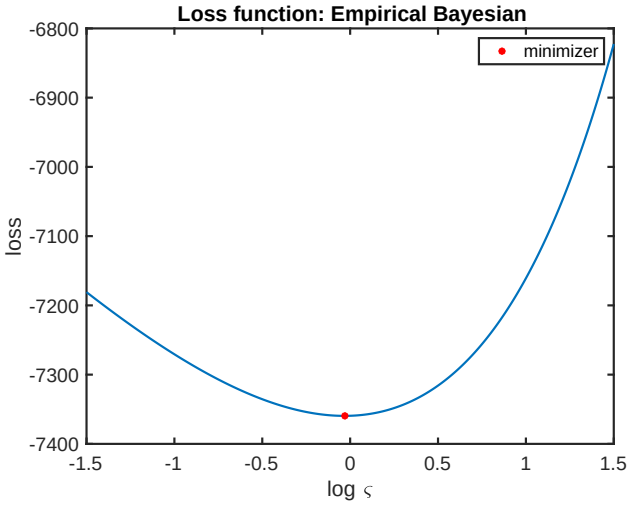

FIGURE 4. EB loss function for recovering σ

to be maximized over to determine our estimate of σ. In our experiments we work with the parameterization $\varsigma = \exp(\varsigma')$ in order to ensure that the estimated σ is positive. Hence, the x-axis of Figure 4 is ς'. The figure shows that the minimizer of the loss function is close to the point $\varsigma' = 0$ ($\varsigma = 1$), so the estimator σ^{EB} is close to the ground truth σ.

We can theoretically analyze the convergence. The same set-up in Subsection 2.1 is adopted, except now we assume the function is drawn from $\mathcal{N}(0, \sigma^2(-\Delta)^{-s})$ with s known and we want to recover σ by seeing the equidistributed spatial samples on the torus. After calculating the likelihood in such a case, we get the EB estimator below. Here we abuse the notation to write

$$\begin{aligned} \sigma^{\mathrm{EB}}(q, u^\dagger) &= \operatorname*{argmin}_{\varsigma>0} \mathsf{L}^{\mathrm{EB}}(\varsigma, q, u^\dagger), \\ \mathsf{L}^{\mathrm{EB}}(\varsigma, q, u^\dagger) &:= \frac{\sigma^2 \|u(\cdot, s, q)\|_s^2}{\varsigma^2} + \log\det K(s, q) + 2^{qd}\log\varsigma^2 \,. \end{aligned} \tag{3.1}$$

The definition of $u(\cdot, s, q), K(s, q)$ is the same as in Subsection 2.1. Recall that $u(\cdot, s, q)$ is the mean of the GP found by conditioning a prior measure $\mathcal{N}(0, (-\Delta)^{-s})$ on observations of $u^\dagger$ at the observation data with level q. The definition of $\|\cdot\|_s$ also follows from Subsection 2.1. We abuse notation to write $\mathsf{L}^{\mathrm{EB}}(\varsigma, q, u^\dagger)$ for the EB loss function used in the estimation of σ; the reader should not confuse this with $\mathsf{L}^{\mathrm{EB}}(t, q, u^\dagger)$ in Subsection 2.1 which is used for recovering the regularity parameter s.

In this setting we have the following consistency result:

Theorem 3.1. *Fix $\delta > 0$. Suppose $u^\dagger$ is a sample drawn from the Gaussian process $\mathcal{N}(0, \sigma^2(-\Delta)^{-s})$ for some $s \in [d/2 + \delta, 1/\delta]$. Then, for the Empirical Bayesian estimator of σ, it holds that*

$$\lim_{q\to\infty} \sigma^{\mathrm{EB}}(q, u^\dagger) = \sigma \,,$$

where the convergence is in probability with respect to randomly chosen $u^\dagger$.

Proof. By taking the derivative of $\mathsf{L}^{\mathrm{EB}}(\varsigma, q, u^\dagger)$ with respect to ς and setting it to 0, we get the explicit formula:

$$\sigma^{\mathrm{EB}}(q, u^\dagger) = \sigma\sqrt{\frac{\|u(\cdot, s, q)\|_s^2}{2^{qd}}} \,. \tag{3.2}$$

Due to Proposition 2.6, we get our $\|u(\cdot, s, q)\|_s^2 = \sum_{m\in B_q^d} \xi_m^2$. By the Law of Large Numbers, we have

$$\lim_{q\to\infty} \frac{\|u(\cdot, s, q)\|_s^2}{2^{qd}} = 1 \,,$$

from which the consistency follows. □

3.1.2. *Recovery of s, σ simultaneously.* We now build on the previous experiment to study whether the EB method can recover s, σ simultaneously when τ is fixed. We reemphasize that since the measures are mutually singular with respect to changes in σ and s we do expect to be able to recover (σ, s) by EB. The basic set-up is the same as the last subsection, and now we minimize the EB loss function to recover s, σ where, again, $\sigma = \exp(\sigma')$. We run 50 instances (each instance corresponds to a random draw of ξ), and collect the estimators $(s^{\mathrm{EB}}, \log\sigma^{\mathrm{EB}})$ of the EB loss function for each instance. We present the histogram of the two values obtained in the experiments as follows (Figure 5).

From the figure, we observe that in the 50 runs, the minimizer $(s^{\mathrm{EB}}, \sigma^{\mathrm{EB}})$ is close to the ground truth $(2.5, 1)$. We conclude that the EB method can recover the two parameters simultaneously in such a context.

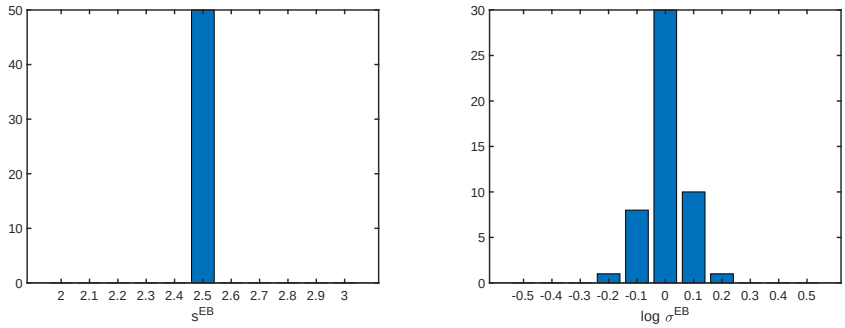

FIGURE 5. Left: histogram of the s^{EB}; right: histogram of the $\log \sigma^{\mathrm{EB}}$

3.1.3. *Recovery of τ.* We consider whether EB and KF can recover the inverse lengthscale parameter τ. We assume that σ is fixed at 1, s is chosen to be 2.5, and sample $u^\dagger \sim \mathcal{N}(0, (-\Delta + \tau^2 I)^{-s})$ with $\tau = 1$. As in the preceding experiments we consider the one dimensional torus example, and the same discretization precision and data acquisition setting as before. We draw 50 instances of $u^\dagger$, and for each of them, calculate the minimizers of the EB and KF loss function. We write $\tau = \exp(\tau')$ and the estimator is $\log \tau^{\mathrm{EB}}$ for τ', which we constrain to be in the interval $[-2, 2]$. In the EB loss function we fix $t = s$ within the loss function; for the KF method, we select $t = s$ (case 1) and $t = \frac{s-d/2}{2}$ (case 2) respectively within the loss function. The histogram of the minimizers of the resulting EB loss function and KF loss functions (in both cases) are presented in Figure 6, expressed in terms of $\log \tau^{\mathrm{EB}}$ and $\log \tau^{\mathrm{KF}}$. In the 50 runs, the EB estimator takes many different values with no apparent pattern. For both case 1 and case 2, the KF estimator of τ' takes the value 2 very often, which is the maximal value of the constrained decision variable. None of the estimators recover the true $\tau' = 0$.

The behavior of the KF estimator can be explained by the observation that when τ increases, the function drawn from the Gaussian prior becomes smoother, and hence the subsampling step in the KF loss does not sacrifice too much information. Therefore, the KF loss exhibits a tendency to get smaller as τ increases. We can understand why EB cannot recover τ by studying the equivalence of Gaussian measures. As shown in [8], when dimension $d \leq 3$, the Gaussian measures $\mathcal{N}(0, (-\Delta + \tau^2 I)^{-s})$ for different τ are equivalent; thus one cannot expect to recover τ using the information from one sample.

We can also consider the problem of recovering s, τ simultaneously, i.e., we solve a joint minimization problem to get $s^{\mathrm{EB}}, \log \tau^{\mathrm{EB}}$ and $s^{\mathrm{KF}}, \log \tau^{\mathrm{KF}}$. The set-up is the same as above, with the sample drawn from $\mathcal{N}(0, (-\Delta + \tau^2 I)^{-s})$ for $\tau = 1$ and $s = 2.5$. We form the EB and KF loss for 50 instances of different draws and find the minimizers as corresponding estimators. The histograms of the estimators are shown in Figure 7 and 8. These figures show that in this joint optimization, the EB method picks the correct value $s^{\mathrm{EB}} = 2.5$ for estimating s, and exhibit no patterns for $\log \tau^{\mathrm{EB}}$; the KF method finds values close to 1 for s^{KF}, as it would in the absence of simultaneous estimation of τ', and selects the largest possible value in the constraint for $\log \tau^{\mathrm{KF}}$, here being 2. The conclusion is that the fact that τ' cannot be learned accurately does not influence the estimation of the regularity

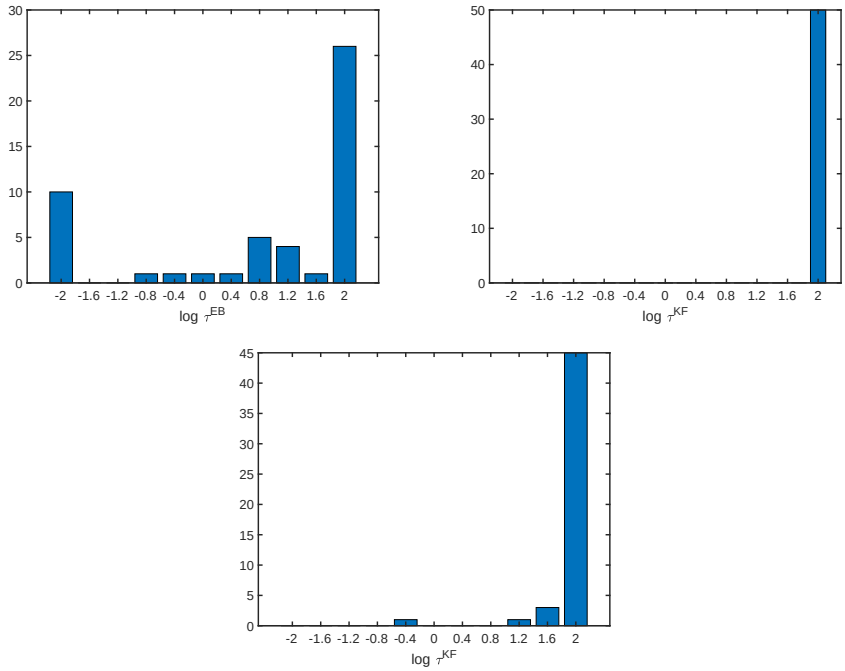

FIGURE 6. Histogram of the $\log \tau^{\mathrm{EB}}$ or $\log \tau^{\mathrm{KF}}$. Upper left: EB loss; upper right: KF loss (case 1); bottom: KF loss (case 2)

parameter s in a context in which the two are learned simultaneously. Indeed, this conclusion also holds when we are recovering the three parameters (s, σ, τ) simultaneously.

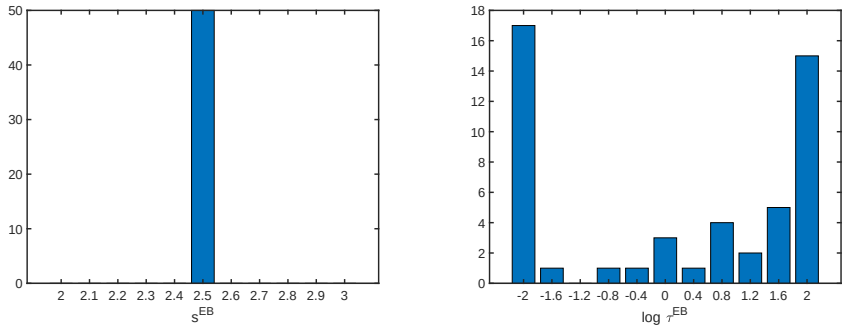

FIGURE 7. EB approach. Left: histogram of the s^{EB}; right: histogram of the $\log \tau^{\mathrm{EB}}$

3.2. **Variance of Regularity Parameter Estimation.** In this subsection, we compare the variance of the two estimators for recovering the regularity parameter s. We return to the experimental set-up in Subsection 2.2.2. We form the EB and KF estimators for 50 instances of different draws of the Gaussian Process,

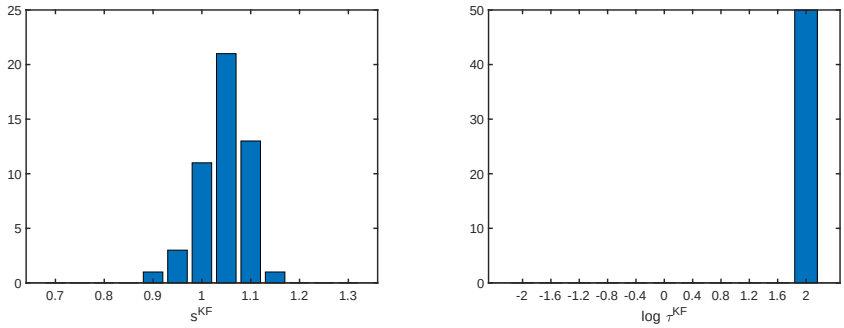

FIGURE 8. KF approach. Left: histogram of s^{KF}; right: histogram of the $\log \tau^{\mathrm{KF}}$

normalized by the limiting optimum values s and $\frac{s-d/2}{2}$ respectively. The statistics of the two estimators are summarized in the histogram (see Figure 9). Clearly, EB

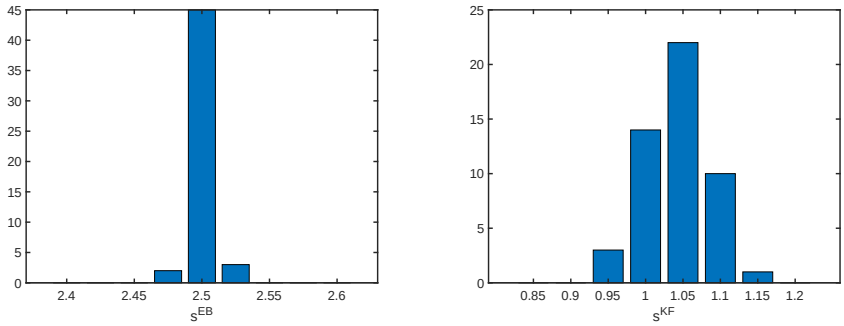

FIGURE 9. Histogram of the regularity estimators for the Matérn-like process. Left: EB; right: KF

exhibits smaller variance than KF. We compute the estimated variance using the 50 instances. Finally we get

$$\frac{\mathrm{Var}(s^{\mathrm{EB}})}{s^2} \approx 1.44 \times 10^{-5} \quad \text{and} \quad \frac{\mathrm{Var}(s^{\mathrm{KF}})}{((s-d/2)/2)^2} \approx 3.6 \times 10^{-3}\,.$$

Since the variance of EB is smaller, if our target is to estimate s for the exact GP model, then this suggests that the EB method is preferable.

3.3. **Other Well-specified Examples.** In this subsection, we consider numerical examples for recovering parameters of a random field in the well-specified case, going beyond the Matérn process studied thus far.

3.3.1. *Recovery of regularity parameter for variable coefficient elliptic operator.* Set $D = [0, 1]$ so that $d = 1$. The theoretical result in Section 2 assumes the function observed $u^\dagger$ is drawn from $\mathcal{N}(0, (-\Delta)^{-s})$ on a torus. In this subsection, we assume $u^\dagger$ is drawn from $\mathcal{N}(0, (-\nabla \cdot (a\nabla \cdot))^{-s})$ for some non-constant function a, and that the elliptic operator implicit in this defintion of a Gaussian measure is equipped

with homogeneous Dirichlet boundary condition on D. We observe its values on the 2^9 equidistributed points of the total 2^{10} grid points used for discretization.

Here we select a coefficient $a(x)$ that exhibits a discontinuity at $x = 1/2$:

$$a(x) = \begin{cases} 1 & x \in [0, 1/2] \\ 2 & x \in (1/2, 1] \end{cases}. \tag{3.3}$$

As a consequence the induced operator is not the Laplacian. We pick $s = 2.5$ to draw a sample $u^\dagger$.

In the well-specified case, the GP used in defining the EB and KF estimators is parameterized by $\mathcal{N}(0, (-\nabla \cdot (a\nabla \cdot))^{-t})$ and we aim to learn parameter t given a data calculated using a draw from the same measure with $t = s$. We consider the well-specified case here (the misspecified case will be considered in Subsection 3.4.1.) We output the histogram of the EB and KF estimators for 50 different draws of $u^\dagger$ in Figure 10. The experiments show that for the variable coefficient elliptic

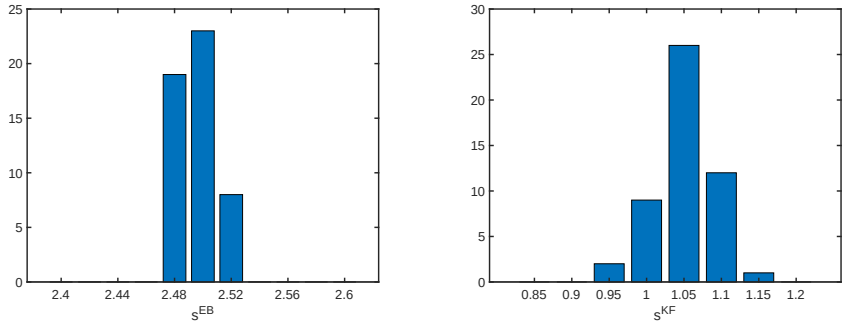

FIGURE 10. Histogram of the regularity estimators for the variable coefficient covariance case. Left: EB; right: KF

operator model, EB and KF succeed in converging to the correct limits. We can calculate the (normalized) variance of the two estimators based on the histograms:

$$\frac{\mathrm{Var}(s^{\mathrm{EB}})}{s^2} \approx 7.8 \times 10^{-5} \quad \text{and} \quad \frac{\mathrm{Var}(s^{\mathrm{KF}})}{((s - d/2)/2)^2} \approx 4 \times 10^{-3}\,.$$

The relative magnitude is similar to the one in Subsection 3.2.

3.3.2. *Recovery of discontinuity position for conductivity field.* Define the conductivity field $a_\theta : [0, 1] \mapsto \mathbb{R}$, and parameterized by $\theta \in [0, 1]$, via

$$a_\theta(x) = \begin{cases} 1 & x \in [0, \theta] \\ 2 & x \in (\theta, 1] \end{cases}. \tag{3.4}$$

In this subsection, we assume that our data $u^\dagger$ is obtained by solving the SPDE

$$-\nabla \cdot (a_{1/2} \nabla u^\dagger) = \xi\,,$$

subject to a homogeneous Dirichlet boundary condition on $[0, 1]$. We choose ξ as a random draw from $\mathcal{N}(0, (-\Delta)^{-1})$. We can view $u^\dagger$ is a sample drawn from $\mathcal{N}(0, C_a)$ where

$$C_a = (-\nabla \cdot (a_{1/2}\nabla \cdot))^{-1}(-\Delta)^{-1}(-\nabla \cdot (a_{1/2}\nabla \cdot))^{-1}. \tag{3.5}$$

We observe the value of $u^\dagger$ on the 2^9 equidistributed points of the total 2^{10} grid points used for discretization. We use EB and KF to estimate θ from the partial observation of the function $u^\dagger$ based on the GP model $\mathcal{N}(0, C_{a,s})$ where

$$C_{a,s} = (-\nabla \cdot (a_\theta \nabla \cdot))^{-1} (-\Delta)^{-s} (-\nabla \cdot (a_\theta \nabla \cdot))^{-1}. \tag{3.6}$$

The model is well-specified for $s = 1$ and misspecified for $s \neq 1$. Here consider the well-specified case in this subsection, i.e., $s = 1$, and $C_{a,s} = C_a$; the misspecified case is covered in Subsection 3.4.2.

We let the domain for θ be $[0.3, 0.7]$ in the definition of EB and KF estimators. We compute the estimators for 50 different draws of $u^\dagger$. The histograms of the EB and KF estimators are shown in Figure 11. The loss functions for one random instance are shown in Figure 12.

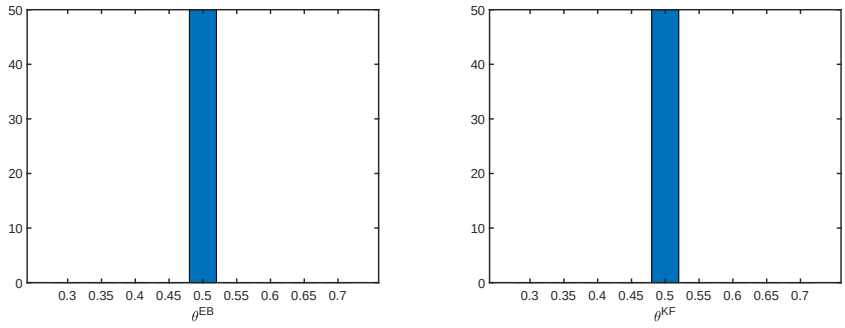

FIGURE 11. Histogram of the discontinuity position estimators (well-specified). Left: EB; right: KF

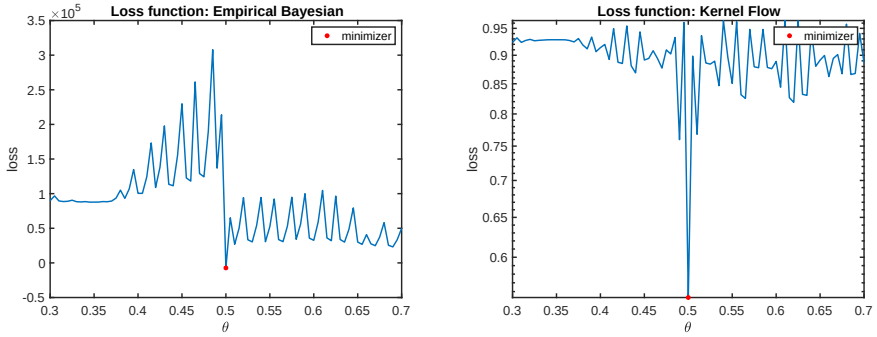

FIGURE 12. Loss function for recovering the discontinuity (well-specified). Left: EB; right: KF

Our experiments show that both EB and KF can recover $\theta = 1/2$, and the recovery is very stable with respect to different draws of $u^\dagger$ from the SPDE. We conclude that the EB and KF can go beyond the Matérn-like kernel model in practice; recovering the point of discontinuity of the conductivity field is an example of this fact.

3.4. **Model Misspecification.** All our preceding experiments are focused on the well-specified case: the function $u^\dagger$ is drawn from the GP model assumed in the estimation, or equivalently, the model for $u^\dagger$ and for the kernel family K_θ in defining the loss functions are matched. This subsection studies model misspecification. We consider two possible ways to misspecify the model: (1) the function $u^\dagger$ is drawn from a GP which is different from that used in defining the loss function; (2) the function $u^\dagger$ is a fixed deterministic function. The second case may arise, for example, if the function comes from a solution of a PDE with some physical data, and there is no natural stochastic context for its provenance. The aim of this subsection is to study the behavior of the EB and KF estimators to compare their robustness to model misspecification.

3.4.1. *Stochastic model misspecification for recovering regularity.* In this subsection, we assume $u^\dagger$ is drawn from $\mathcal{N}(0, (-\nabla \cdot (a\nabla \cdot))^{-s})$, while the GP used in defining the EB and KF estimators is still $\mathcal{N}(0, (-\Delta)^{-t})$. This results in a model misspecification corresponding to the well-specified model in Subsection 3.3.1. As in Subsubection 3.3.1, we select a as in (3.3) and we set $s = 2.5$ to draw the sample $u^\dagger$. Figure 13 shows the histograms of the minimizers of the EB and KF loss functions obtained from 50 independent draws from the Gaussian Process. Despite misspecification, the EB and KF estimators are still concentrated around 2.5 and 1, respectively. We also calculate the variance:

$$\frac{\mathrm{Var}(s^{\mathrm{EB}})}{s^2} \approx 5.9 \times 10^{-4} \quad \text{and} \quad \frac{\mathrm{Var}(s^{\mathrm{KF}})}{((s-d/2)/2)^2} \approx 6.8 \times 10^{-4}\,.$$

In this example, the (normalized) variance of KF of EB are of similar magnitude. This is different from the well-specified case in Subsection 3.3.1 where the variance of EB is much smaller than KF.

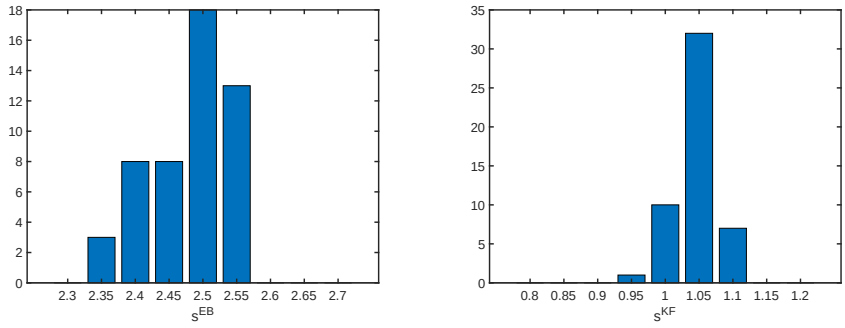

FIGURE 13. Histogram of the regularity estimators under model misspecification. Left: EB; right: KF

3.4.2. *Stochastic model misspecification for recovering discontinuity.* In this subsection, we consider the model misspecification corresponding to the well-specified case in Subsection 3.3.2. For the GP defining the EB and KF estimators we use the centred Gaussian with covariance operator given by (3.6) with $s = 5$; meanwhile $u^\dagger$ is drawn from the centred Gaussian with covariance operator given by (3.5); thus we are in a misspecified version of the setting arising in Subsection 3.3.2 and,

as there, our aim is to recover the point of discontinuity. We illustrate the loss functions for a single draw of $u^\dagger$ in Figure 14. These plots are not sensitive to the particular draw of $u^\dagger$ and illustrate the robustness of KF (and the lack of robustness of EB) to this misspecification. Indeed, the EB estimator gives 0.3 which is the lower boundary of the compact parameter space used in the minmization, while the KF estimator picks the true parameter 0.5. The loss function of KF, shown in Figure 14, exhibits a sharp global minimizer at $\theta = 0.5$.

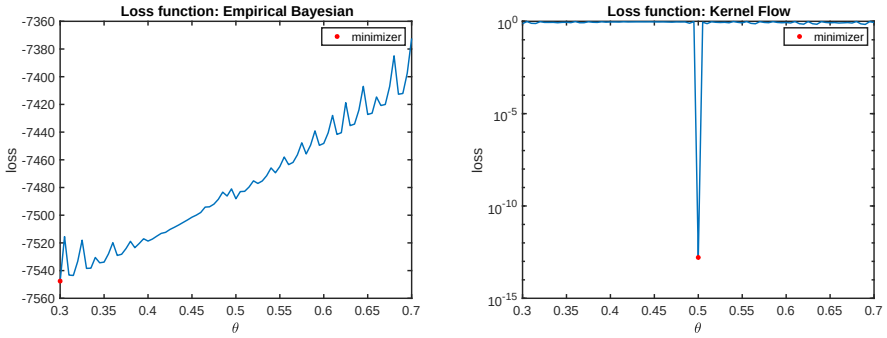

FIGURE 14. Loss function for estimating the discontinuity parameter under model misspecification. Left: EB; right: KF

3.4.3. *Deterministic model.* In this subsection, we consider the EB and KF estimators for the parameter t in the GP model $\mathcal{N}(0, (-\Delta)^{-t})$ where Δ is equipped with homogeneous Dirichlet boundary conditions on $[0, 1]$. However, rather than choosing $u^\dagger$ that is drawn from the GP $\mathcal{N}(0, (-\Delta)^{-s})$ for some s (as we did in Section 2), we choose it be the solution to the equation $(-\Delta)^s u^\dagger(\cdot) = \delta(\cdot - 1/2)$, i.e., $u^\dagger$ is the Green function corresponding to the differential operator $(-\Delta)^s$ and evaluated at $y = 1/2$. Since $u^\dagger$ has no stochastic background, we understand this situation as a deterministic model misspecification.

We observe the value of $u^\dagger$ on the 2^9 equidistributed points of the total 2^{10} grid points used for discretization. We conduct numerical experiments to find the value of the EB and KF estimators. Our experiments show that the EB estimator returns $2s$ and the KF estimator returns s for this one dimensional example. The loss function in the case $s = 1.2$ is shown in Figure 15.

We now describe some regularity considerations in order to understand the observed phenomenon. In this one dimensional example, $\delta(\cdot - 1/2)$ belongs to $H^\eta([0, 1])$ for any $\eta < -1/2$, so the solution $u \in H^{2s+\eta}([0, 1])$ for any $\eta < -1/2$. It is of critical regularity $2s - 1/2$, but this criticality is not homogeneous: it is caused by the presence of a singularity induced by the Dirac function.

The discussion in Section 2 implies KF will recover $s - 1/4$ while EB recovers $2s$ for a function with homogeneous critical regularity $2s - 1/2$. However, the experiments here show that KF recovers s while EB recovers $2s$, for this function with critical regularity $2s-1/2$; unlike the setting in Section 2, here the ground truth lacks spatial homogeneity. This suggests that the KF estimator for the regularity parameter is sensitive to whether the regularity of the target function is spatially homogeneous or not. This fact is not surprising, considering the vast literature on adaptive approximation for functions with singularities, which implies the presence

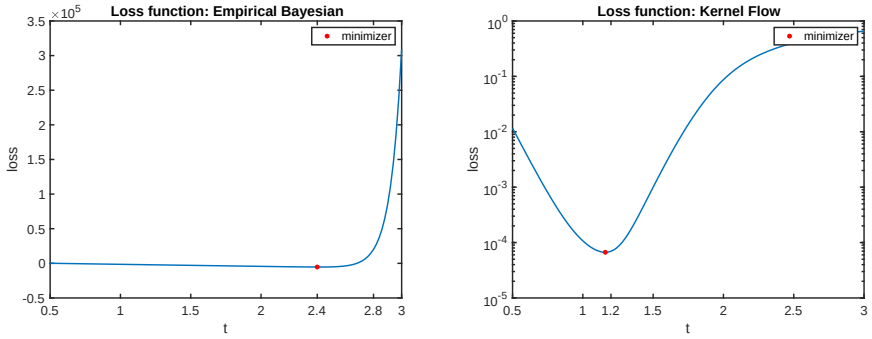

FIGURE 15. Loss function for estimating the regularity parameter under deterministic $u^\dagger$. Left: EB; right: KF

of a singularity will exert considerable influence on the approximation error resulting from minimizing the KF loss function. In this example, the optimal approximation in KF error comes at $t = s$. We can understand this phenomenon as follows. Recall $u^\dagger = (-\Delta)^{-s}\delta(\cdot - 1/2)$. Using $\mathcal{N}(0, (-\Delta)^{-t})$ in the GPR is equivalent to using the basis functions $\text{span}_{j\in J_q}\{(-\Delta)^{-t}\delta(\cdot - x_j)\}$ (as in Section 2.1) with x_i being the data points indexed by $j \in J_q$, to approximate $u^\dagger$. When $t = s$ and one of the $x_j = 1/2$, the ground truth will just be in the basis functions set, so it is straightforward to imagine $t = s$ leads to the smallest approximation error, and KF picks this value.

We understand the fact that EB still picks $t = 2s$ by making the following observation: there are only two terms in the EB loss function. The log determinant term remains the same for each t when $u^\dagger$ changes. For the norm term $\|u(\cdot, t, q)\|_t^2$, the blow-up rate depends on the regularity of $u^\dagger$. Here, it makes no difference whether the regularity of $u^\dagger$ is spatially homogeneous or not.

3.5. **Computational Aspects.** Practical applications require weighting statistical efficiency against computational complexity. Although the regularity models covered in this paper appear to produce well-behaved EB and KF loss functions with easily identifiable global minimizers, models with high dimensional parameter space typically require using algorithms such as gradient descent which do not come with theoretical guarantees on the identification of global minimizers. Furthermore, when the size of the data is large, computation becomes a limiting factor, and subsampling offers a traditional remedy when combined with gradient descent, but again theoretical guarantees are not typically to be expected. The stochastic algorithm presented in [20] for KF can be interpreted as an SGD algorithm aimed at minimizing the average loss

$$\mathbb{E}_{\pi_1}\mathbb{E}_{\pi_2}\mathsf{L}^{\mathrm{KF}}(\theta, \pi_1\mathcal{X}, \pi_2\pi_1\mathcal{X}, u^\dagger)\,,$$

via draws from the distribution of π_1 and π_2 ($\pi_1\mathcal{X}$ is a random subsampling of $\mathcal{X}$, and $\pi_2\pi_1\mathcal{X}$ is a further random subsampling of $\pi_1\mathcal{X}$). The efficacy of an analogous strategy for EB remains unclear due to the presence of the log determinant term in the loss. It is of future interest to explore further the computational aspects of the EB and KF approaches to hierarchical learning.

4. Discussions

In this paper, we have studied the Empirical Bayes and Kernel Flow approaches to hyperparameter learning. The first approach is based on statistical considerations, while the second approach originates from an approximation theoretic viewpoint. Their distinct objectives lead them to different behaviors and different interpretations of optimality.

For the Matérn-like process model, we made a detailed theoretical study of the recovery of the regularity parameter. We proved the EB estimator converges to s, while the KF estimator converges to $\frac{s-d/2}{2}$, both results holding in probability in the large data limit if the regularity of the GP that $u^\dagger$ draws from is s. Our experiments illustrate that, in terms of the L^2 error $\|u(\cdot,t,q)-u^\dagger\|_0^2$, the parameter $t=\frac{s-d/2}{2}$ relates to the minimal t that achieves the fast error rate while $t=s$ relates to the t that achieves the smallest error, averaged over the GP $u^\dagger\sim\mathcal{N}(0,(-\Delta)^{-s})$. This demonstrates the different drivers that guide the EB and KF methods in selecting the parameters. The statistical and approximation theoretic principles behind them lead to the differences between them.

In the theoretical study, we developed a Fourier analysis toolkit for this problem, and as a byproduct, we showed the consistency of recovering σ in the Matérn-like process for the EB method. Recovery of the lengthscale parameter and recovery of several parameters simultaneously was studied via numerical experiments. It is of future interest to perform theoretical studies explaining these empirically observed phenomena. Furthermore, the theory in this paper is based on an equidistributed design for the data location, and the generalization to randomized design remains a potential further direction. Also, our focus in this paper is on the noiseless observation setting, and an extension to the noisy case is of future theoretical interest.

Our numerical experiments for additional well-specified and misspecified models extend the scope of this paper beyond the Matérn-like kernels. Both the two estimators work very well in the well-specified models we consider; we would like to explore this more in the future, both theoretically and numerically, potentially in more complex models that are present in machine learning. The variance and robustness of the estimators behave differently for the misspecified models. The variabilities in robustness are in line with our expectation since these estimators follow from different decision rules; these rules can vary considerably in sensitivity to model mismatches of different kinds. In practice, users should choose the correct approach to avoid high sensitivity to likely model errors present.

As a summary, this paper demonstrates some basic aspects of the difference between Bayesian and approximation theoretic approaches for hierarchical learning. Generally, it is of interest to study EB and KF for other types of models and to study other parameter selection criteria based on the two principles beyond EB and KF, such as a fully Bayesian approach or another choice of d for the approximation, and identify their pros and cons under different scenarios. We are interested in exploring the theoretical and practical performance of methods under such a framework, and we believe that a diversity in such methods will enable users to deal with the model misspecification that is to be expected in many applications.

Acknowledgements YC gratefully acknowledges the support of the Caltech Kortchack Scholar Program. HO gratefully acknowledges support from AFOSR (grant FA9550-18-1-0271) and ONR (grant N00014-18-1-2363). AMS is grateful to AFOSR (grant FA9550-17-1-0185) and NSF (grant DMS 18189770) for financial support.

References

1. D. M. Allen. The relationship between variable selection and data agumentation and a method for prediction. *technometrics*, 16(1):125–127, 1974.
2. S.-i. Amari and S. Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6):783–789, 1999.
3. F. Bachoc. Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics & Data Analysis*, 66:55–69, 2013.
4. F. Bachoc, A. Lagnoux, and T. M. N. Nguyen. Cross-validation estimation of covariance parameters under fixed-domain asymptotics. *Journal of Multivariate Analysis*, 160:42–67, 2017.
5. C. Cortes, M. Kloft, and M. Mohri. Learning kernels using local rademacher complexity. In *Advances in neural information processing systems*, pages 2760–2768, 2013.
6. M. Dashti and A. M. Stuart. The Bayesian approach to inverse problems. *arXiv preprint arXiv:1302.6989*, 2013.
7. C. De Boor, R. A. DeVore, and A. Ron. Approximation from shift-invariant subspaces of $l^2(\mathbb{R}^d)$. *Transactions of the American Mathematical Society*, 341(2):787–806, 1994.
8. M. M. Dunlop, M. A. Iglesias, and A. M. Stuart. Hierarchical bayesian level set inversion. *Statistics and Computing*, 27(6):1555–1584, 2017.
9. S. Geisser. The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328, 1975.
10. J. K. Ghosh and R. Ramamoorthi. *Bayesian Nonparametrics*. Springer Science & Business Media, 2003.
11. N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. *Bayesian Nonparametrics*, volume 28. Cambridge University Press, 2010.
12. T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, pages 1171–1220, 2008.
13. A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.
14. B. T. Knapik, B. Szabó, A. W. Van Der Vaart, and J. Van Zanten. Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probability Theory and Related Fields*, 164(3-4):771–813, 2016.
15. B. T. Knapik, A. W. Van Der Vaart, J. H. van Zanten, et al. Bayesian inverse problems with Gaussian priors. *The Annals of Statistics*, 39(5):2626–2657, 2011.
16. R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.
17. R. Kohn, C. F. Ansley, and D. Tharm. The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *Journal of the american statistical association*, 86(416):1042–1050, 1991.
18. J. Lee, L. Xiao, S. S. Schoenholz, Y. Bahri, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.
19. H. Owhadi and C. Scovel. *Operator-Adapted Wavelets, Fast Solvers, and Numerical Homogenization: From a Game Theoretic Approach to Numerical Approximation and Algorithm Design*, volume 35. Cambridge University Press, 2019.
20. H. Owhadi and G. R. Yoo. Kernel flows: From learning kernels from data into the abyss. *Journal of Computational Physics*, 389:22–47, 2019.
21. C. E. Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

22. A. Ron. The l^2-Approximation Orders of Principal Shift-Invariant Spaces Generated by a Radial Basis Function. In *Numerical Methods in Approximation Theory, Vol. 9*, pages 245–268. Birkhuser Basel, Basel, 1992.
23. M. Scheuerer, R. Schaback, and M. Schlather. Interpolation of spatial data–a stochastic or a deterministic problem? *European Journal of Applied Mathematics*, 24(4):601–629, 2013.
24. M. L. Stein. A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. *The Annals of Statistics*, pages 1139–1157, 1990.
25. C. J. Stone et al. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12(4):1285–1297, 1984.
26. A. Stuart and A. Teckentrup. Posterior consistency for Gaussian process approximations of Bayesian posterior distributions. *Mathematics of Computation*, 87(310):721–753, 2018.
27. A. L. Teckentrup. Convergence of Gaussian process regression with estimated hyper-parameters and applications in Bayesian inverse problems. *arXiv preprint arXiv:1909.00232*, 2019.
28. M. J. van der Laan, S. Dudoit, and A. W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics & Decisions*, 24(3):373–395, 2006.
29. A. van der Vaart and J. A. Wellner. *Weak Convergence And Empirical Processes*. 1996.
30. A. W. Van der Vaart, S. Dudoit, and M. J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24(3):351–371, 2006.
31. G. Wahba and J. Wendelberger. Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly weather review*, 108(8):1122–1143, 1980.
32. J. Warnes and B. Ripley. Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, 74(3):640–642, 1987.
33. H. Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.
34. A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Deep kernel learning. In *Artificial Intelligence and Statistics*, pages 370–378, 2016.
35. Y. Yang et al. Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473, 2007.
36. Z. Ying. Asymptotic properties of a maximum likelihood estimator with data from a gaussian process. *Journal of Multivariate Analysis*, 36(2):280–296, 1991.
37. G. R. Yoo and H. Owhadi. Deep regularization and direct training of the inner layers of neural networks with kernel flows. *arXiv preprint arXiv:2002.08335*, 2020.
38. H. Zhang, Y. Wang, et al. Kriging and cross-validation for massive spatial data. *Environmetrics*, 21(3/4):290–304, 2010.

5. Appendix: Proofs

5.1. **Proof of Proposition 2.3.**

Proof. Let $\varphi_j(x) = (-\Delta)^{-t}\delta(x - x_j)$ and in particular $\varphi_0(x) = (-\Delta)^{-t}\delta(x)$. We have for $m \in \mathbb{Z}^d$,

$$\hat{\varphi}_0(m) = \begin{cases} (4\pi^2)^{-t}|m|^{-2t}, & m \neq 0 \\ 0, & m = 0\,. \end{cases}$$

We introduce the translation operator $\tau_{j2^{-q}}$ which acts on function $u : \mathbb{T}^d \to \mathbb{R}$ and is defined by

$$(\tau_{j2^{-q}}u)(x) = u(x_1 - j_1 2^{-q}, x_2 - j_2 2^{-q}, ..., x_d - j_d 2^{-q})$$

for $j = (j_1, j_2, ..., j_d) \in \mathbb{Z}^d$ and $x = (x_1, x_2, ..., x_d) \in \mathbb{R}^d$. Then, for $j \in J_q$, we have the relation $\delta(\cdot - x_j) = \tau_{j2^{-q}}\delta(\cdot)$. Using the property of the Fourier coefficients, we obtain

$$\hat{\varphi}_j(m) = \hat{\varphi}_0(m)e^{-2\pi i\langle j2^{-q},m\rangle} = \begin{cases} (4\pi^2)^{-t}|m|^{-2t}e^{-2\pi i\langle j2^{-q},m\rangle}, & m \neq 0 \\ 0, & m = 0\,. \end{cases}$$

By definition, $\hat{\mathcal{F}}_{t,q}$ is the span of such $\hat{\varphi}_j$ for $j \in J_q$. Hence, for any $g \in \hat{\mathcal{F}}_{t,q}$, it can be written as a linear combination of these functions. Equivalently, there exists a 2^q-periodic function p such that

$$g(m) = \begin{cases} |m|^{-2t} p(m), & m \neq 0 \\ 0, & m = 0\,. \end{cases}$$

This gives the desired representation of g. □

5.2. **Proof of Theorem 2.5.**

Proof. By Proposition 2.3, there exists a 2^q-periodic function $p_1(m)$ on $\mathbb{Z}^d$, such that,

$$\hat{u}(m,t,q) = \begin{cases} |m|^{-2t} p_1(m), & m \neq 0 \\ 0, & m = 0\,. \end{cases}$$

By the definition of GPR, we have $[u^\dagger(\cdot) - u(\cdot,t,q), \delta(\cdot - x_j)] = 0$ for every data point x_j. In the Fourier domain, according to the characterization of $\hat{\mathcal{F}}_{t,q}$, this orthogonality leads to

$$\sum_{m \in \mathbb{Z}^d} (\hat{u}(m) - \hat{u}(m,t,q)) p(m) = 0 \tag{5.1}$$

for $p : \mathbb{Z}^d \to \mathbb{C}$ being any 2^q-periodic function. Recalling Definition 2.4, we have

$$(T_q \hat{u})(m) = \sum_{\beta \in \mathbb{Z}^d} \hat{u}(m + 2^q \beta)\,. \tag{5.2}$$

The fact that the above sum converges may be seen as a consequence of the CauchySchwarz inequality and the regularity of u (recall $t \geq d/2 + \delta$). Using (5.2) and the representation of $\hat{u}(m,t,q)$, we reformulate (5.1) as

$$\sum_{m \in B_q^d} \left((T_q \hat{u})(m) - M_q^t(m) p_1(m) \right) p(m) = 0\,.$$

The above formula holds for any 2^q-periodic function p. Let $g(m) = (T_q \hat{u})(m) - M_q^t(m) p_1(m)$, then we get that g is a 2^q-periodic function on $\mathbb{Z}^d$ and that

$$\sum_{m \in B_q^d} g(m) p(m) = 0$$

holds for any 2^q-periodic function p. This implies that $g(m) = 0$. Hence, we get

$$p_1(m) = \frac{(T_q \hat{u})(m)}{M_q^t(m)}\,.$$

Plugging this expression into the above representation formula for $\hat{u}(m,t,q)$ leads to

$$\hat{u}(m,t,q) = \begin{cases} 0, & \text{if } m = 0 \\ |m|^{-2t} \frac{(T_q \hat{u})(m)}{M_q^t(m)}, & \text{else}\,. \end{cases}$$

This completes the proof. □

5.3. **Proof of Lemma 2.7.**

Proof. Recall the definition

$$M_q^t(m) := \begin{cases} \sum_{\beta\in\mathbb{Z}^d\setminus\{0\}} |2^q\beta|^{-2t}, & \text{if } m = j\cdot 2^q \text{ for some } j\in\mathbb{Z}^d \\ \sum_{\beta\in\mathbb{Z}^d} |m+2^q\beta|^{-2t}, & \text{else}\,. \end{cases}$$

Because of the periodicity of M_q^t, we need only to study $m\in B_q^d$. We split it into two cases.

(1) If $m=0$, then $M_q^t(m) = \sum_{\beta\in\mathbb{Z}^d\setminus\{0\}} |2^q\beta|^{-2t} = 2^{-2qt}\sum_{\beta\in\mathbb{Z}^d\setminus\{0\}}|\beta|^{-2t} \simeq 2^{-2qt}$.

(2) If $m\in B_q^d\setminus\{0\}$, then $M_q^t(m) = \sum_{\beta\in\mathbb{Z}^d}|m+2^q\beta|^{-2t} = |m|^{-2t} + \sum_{\beta\in\mathbb{Z}^d\setminus\{0\}}|m+2^q\beta|^{-2t}$. Since $B_q^d = [-2^{q-1}, 2^{q-1}-1]^{\otimes d}$, each component of $m\in B_q^d$ is bounded by 2^{q-1} in amplitude, and therefore each component of $2^{-q}m$ is bounded by $1/2$ in amplitude. So, it follows that

$$\sum_{\beta\in\mathbb{Z}^d\setminus\{0\}} |m+2^q\beta|^{-2t} = 2^{-2qt}\sum_{\beta\in\mathbb{Z}^d\setminus\{0\}} |2^{-q}m+\beta|^{-2t} \simeq 2^{-2qt}\,.$$

Then, we get $|m|^{-2t}\le M_q^t(m)\lesssim |m|^{-2t}+2^{-2qt}\lesssim |m|^{-2t}$ where we have used the fact that $|m|\lesssim 2^q$. Therefore, it holds that $M_q^t(m)\simeq |m|^{-2t}$.

As a byproduct of the above proof, we also get $M_q^t(m)-|m|^{-2t}\simeq 2^{-2qt}$. □

5.4. **Proof of Lemma 2.9.**

Proof. First, we prove the pointwise convergence (i.e., for each fixed r), then move on to prove uniform convergence. To achieve this, we calculate the variance:

$$\begin{aligned}\mathrm{Var}(\alpha(r,q)) &\simeq 2^{-2rq}\sum_{m\in B_q^d\setminus\{0\}} |m|^{2r-2d} \\ &\lesssim 2^{-2rq}\int_1^{2^q} x^{2r-2d+d-1}\,\mathrm{d}x = 2^{-2rq}\int_1^{2^q} x^{2r-d-1}\,\mathrm{d}x\,.\end{aligned}$$

For $r=d/2$, the integral gives $\log(2^q)=q\log 2$; for $r\neq d/2$, it is $\frac{1}{2r-d}(2^{q(2r-d)}-1)$. In both cases, we have $\lim_{q\to\infty}\mathrm{Var}(\alpha(r,q))=0$. Thus, $\alpha(r,q)$ converges in L^2 to the limit of its expectation, which we may calculate as follows:

$$\lim_{q\to\infty}\mathbb{E}\alpha(r,q) = \lim_{q\to\infty}\sum_{m\in B_q^d\setminus\{0\}} (2^{-q})^d|2^{-q}m|^{r-d} = \int_{[-1/2,1/2]^d}|y|^{r-d}\,\mathrm{d}y := \gamma(r) > 0\,.$$

Hence, we get $\lim_{q\to\infty}\alpha(r,q)=\gamma(r)>0$ in L^2 for every $r\in[\epsilon,1/\epsilon]$, and the convergence also holds in probability. We may now proceed to show uniform convergence. We rely on Exercise 3.2.3 in [29]. Based on that, it suffices to prove $\alpha(r,q)$ is uniformly Lipschitz continuous as a function of r for $q\in\mathbb{N}$. Pick any $r_1,r_2\in[\epsilon,1/\epsilon]$, then

$$\begin{aligned}&|\alpha(r_1,q)-\alpha(r_2,q)| \\ =&\sum_{m\in B_q^d\setminus\{0\}} 2^{-qd}|(|2^{-q}m|^{r_1-d}-|2^{-q}m|^{r_2-d})| \\ \le&\sum_{m\in B_q^d\setminus\{0\}} 2^{-qd}|r_1-r_2|(|2^{-q}m|^{\epsilon-d}+|2^{-q}m|^{1/\epsilon-d})|\log(2^{-q}m)|\xi_m^2\,,\end{aligned}$$

where in the last step we have used the fact that $||2^{-q}m|^{r_1-d} - |2^{-q}m|^{r_2-d}| = ||2^{-q}m|^{\eta-d}\log(2^{-q}m)(r_1-r_2)|$ for some η that lies between r_1 and r_2, and we use the bound $r_1, r_2 \in [\epsilon, 1/\epsilon]$. Now, we define the random series:

$$\mathsf{L}(q) := 2^{-qd} \sum_{m \in B_q^d \setminus \{0\}} (|2^{-q}m|^{\epsilon-d} + |2^{-q}m|^{1/\epsilon-d})|\log(2^{-q}m)|\xi_m^2 \,.$$

We calculate its variance as follows:

$$\begin{aligned}
\mathrm{Var}(\mathsf{L}(q)) &\simeq 2^{-2dq} \sum_{m \in B_q^d \setminus \{0\}} (|2^{-q}m|^{2\epsilon-2d} + |2^{-q}m|^{2/\epsilon-2d}) \log^2 |2^{-q}m| \\
&\lesssim 2^{-qd} \left(\int_{2^{-q}}^1 t^{2\epsilon-2d+d-1} \log^2 t \, \mathrm{d}t + \int_{2^{-q}}^1 t^{2/\epsilon-2d+d-1} \log^2 t \, \mathrm{d}t \right) \\
&= 2^{-qd} \int_{2^{-q}}^1 (t^{2\epsilon-d-1} + t^{2/\epsilon-d-1}) \log^2 t \, \mathrm{d}t \,, \\
&\lesssim 2^{-qd} \int_{2^{-q}}^1 (t^{\epsilon-d-1} + t^{1/\epsilon-d-1}) \, \mathrm{d}t \lesssim 2^{-q\epsilon} \,.
\end{aligned}$$

The last term will go to 0 as q goes to infinity. Thus, $\mathsf{L}(q)$ converges in L^2 (and thus in probability) to $\mathsf{L}^* = \lim_{q\to\infty} \mathbb{E}\mathsf{L}(q)$, which is

$$\begin{aligned}
\lim_{q\to\infty} \mathbb{E}\mathsf{L}(q) &= \lim_{q\to\infty} 2^{-qd} \sum_{m \in B_q^d \setminus \{0\}} (|2^{-q}m|^{\epsilon-d} + |2^{-q}m|^{1/\epsilon-d}) \log^2 |2^{-q}m| \\
&= \int_{[-1/2,1/2]^d} (|y|^{\epsilon-d} + |y|^{1/\epsilon-d}) \log^2 |y| \, \mathrm{d}y \\
&\lesssim \int_{[-1/2,1/2]^d} (|y|^{\epsilon/2-d} + |y|^{1/(2\epsilon)-d}) \, \mathrm{d}y < \infty \,.
\end{aligned}$$

Using Markov's inequality we deduce that, for any $\epsilon' > 0$, it holds that

$$\mathbb{P}(|\mathsf{L}(q) - \mathsf{L}^*| \geq \epsilon') \leq \frac{\mathbb{E}|\mathsf{L}(q) - \mathsf{L}^*|^2}{(\epsilon')^2} \leq \frac{2^{-q\epsilon}}{(\epsilon')^2} \,.$$

Thus,

$$\sum_{q=1}^{\infty} \mathbb{P}(|\mathsf{L}(q) - \mathsf{L}^*| \geq \epsilon') \leq \sum_{q=1}^{\infty} \frac{2^{-q\epsilon}}{(\epsilon')^2} < \infty \,.$$

From the Borel-Cantelli lemma it follows that $\lim_{q\to\infty} \mathsf{L}(q) = \mathsf{L}^*$ almost surely, and therefore $\mathsf{L}(q)$ is bounded uniformly for q almost surely. Since $|\alpha(r_1, q) - \alpha(r_2, q)| \leq \mathsf{L}(q)|r_1 - r_2|$, it follows that $\alpha(r, q)$ is uniformly Lipschitz continuous as a function of r for $q \in \mathbb{N}$. Invoking Exercise 3.2.3 in [29] concludes this case.

For the case $r = 0$, we follow the same strategy as in the previous case. First, we calculate the corresponding variance:

$$\begin{aligned}
\mathrm{Var}(\alpha(0, q)) &\simeq \frac{1}{q^2} \sum_{m \in B_q^d \setminus \{0\}} |m|^{-2d} \\
&\lesssim \frac{1}{q^2} \int_1^{2^q} x^{-2d+d-1} \, \mathrm{d}x \lesssim \frac{1}{q^2}
\end{aligned}$$

where the last term goes to 0 as q goes to infinity. Then, we calculate the expectation:

$$\mathbb{E}\alpha(0,q) = \frac{1}{q} \sum_{m \in B_q^d \setminus \{0\}} |m|^{-d}.$$

The limit when $q \to \infty$ is identified through the following calculations:

$$\begin{aligned}\lim_{q\to\infty} \frac{1}{q} \sum_{m \in B_q^d \setminus \{0\}} |m|^{-d} &= \lim_{q\to\infty} \sum_{m \in B_{q+1}^d \setminus B_q^d} |m|^{-d} \\ &= \lim_{q\to\infty} 2^{-qd} \sum_{m \in B_{q+1}^d \setminus B_q^d} |2^{-q}m|^{-d} \\ &= \int_{[-1,1]^d \setminus [-1/2,1/2]^d} |x|^{-d}\,\mathrm{d}x < \infty;\end{aligned}$$

here we have used the definition of the Riemann integral. Finally, we conclude that $\lim_{q\to\infty} \alpha(0,q) = \gamma(0)$ in probability for $\gamma(0) \in (0,\infty)$. □

5.5. **Proof of Proposition 2.10.**

Proof. First, we have the relation

$$\log\det K(t,q) = \log\det K(t,q-1) + \log\det(K(t,q)/K(t,q-1))$$

where $K(t,q)/K(t,q-1)$ is the Schur complement of $K(t,q-1)$ in $K(t,q)$. Due to the variational property of the Schur complement (see Lemma 13.24 in [19]), the smallest and largest eigenvalues of $K(t,q)/K(t,q-1)$ satisfy (in the dual norm $\|\cdot\|_{-t}$)

(5.3)

$$\begin{aligned}&\lambda_{\min}(K(t,q)/K(t,q-1)) \geq \inf_{y\in\mathbb{R}^{|J_q|}} \frac{\|\sum_{j\in J_q} y_j \delta(x-x_j)\|_{-t}^2}{|y|^2}, \quad \text{and} \\ &\lambda_{\max}(K(t,q)/K(t,q-1)) \\ &\qquad = \sup_{y\in\mathbb{R}^{|J_q|}} \inf_{z\in\mathbb{R}^{|J_{q-1}|}} \frac{\|\sum_{j\in J_q} y_j \delta(x-x_j) - \sum_{j'\in J_{q-1}} z_{j'}\delta(x-x_{j'})\|_{-t}^2}{|y|^2}.\end{aligned}$$

These two formulae will be crucial in the subsequent analysis. We start by estimating the smallest and largest eigenvalues of the Schur complement. Let $w = (-\Delta)^{-t}\sum_{j\in J_q} y_j\delta(x-x_j)$, whose Fourier coefficients are

$$\hat{w}(m) = \begin{cases} 0, & \text{if } m = 0 \\ (4\pi^2)^{-t}|m|^{-2t} g(m), & \text{else}, \end{cases} \tag{5.4}$$

where, the function $g(m)$ is defined by

$$g(m) = \sum_{j\in J_q} y_j \exp(2\pi i \langle j2^{-q}, m\rangle). \tag{5.5}$$

For the smallest eigenvalue, we write

$$\begin{aligned}\|\sum_{j\in J_q} y_j\delta(x-x_j)\|_{-t}^2 = \|w\|_t^2 &= (4\pi^2)^t \sum_{m\in\mathbb{Z}^d\setminus\{0\}} |m|^{2t}|\hat{w}(m)|^2 \\ &= (4\pi^2)^{-t} \sum_{m\in\mathbb{Z}^d\setminus\{0\}} |m|^{-2t}|g(m)|^2.\end{aligned}$$

Notice that

$$\sum_{m\in\mathbb{Z}^d\setminus\{0\}} |m|^{-2t}|g(m)|^2 = \sum_{m\in B_q^d} M_q^t(m)|g(m)|^2 \gtrsim 2^{-2tq} \sum_{m\in B_q^d} |g(m)|^2$$

and

$$\begin{aligned}\sum_{m\in B_q^d} |g(m)|^2 &= \sum_{m\in B_q^d} |\sum_{j\in J_q} y_j \exp(2\pi i\langle j2^{-q}, m\rangle|^2 \\ &= \sum_{m\in B_q^d}\sum_{j\in J_q}\sum_{l\in J_q} y_j y_l \exp(2\pi i\langle (j-l)2^{-q}, m\rangle \\ &= \sum_{j\in J_q}\sum_{l\in J_q} y_j y_l \sum_{m\in B_q^d} \exp(2\pi i\langle (j-l)2^{-q}, m\rangle \\ &\simeq 2^{qd}|y|^2\,.\end{aligned} \tag{5.6}$$

In the last line we have used the fact that

$$\sum_{m\in B_q^d} \exp(2\pi i\langle (j-l)2^{-q}, m\rangle = \begin{cases} 0, & \text{if } j-l\neq 0 \\ \sum_{m\in B_q^d} 1 \simeq 2^{qd}, & \text{if } j-l=0\,. \end{cases}$$

Thus, combining the above results, we obtain the bound on the smallest eigenvalue

$$\lambda_{\min}(K(t,q)/K(t,q-1)) \gtrsim 2^{-q(2t-d)}\,.$$

We then move to consider the largest eigenvalue. First, notice that

$$\inf_{z\in\mathbb{R}^{|J_{q-1}|}} \|\sum_{j\in J_q} y_j\delta(x-x_j) - \sum_{j'\in J_{q-1}} z_{j'}\delta(x-x_{j'})\|_{-t}^2 = \inf_{v\in\mathcal{F}_{t,q-1}} \|w-v\|_t^2\,.$$

Naturally, one can express the optimal v in the above variational formulation using the Fourier series representation explained before. However, this will lead to many interactions between different frequencies. To make the analysis cleaner, we adopt another strategy. We first approximate the function w by a band-limited function, whose projection into $\mathcal{F}_{t,q-1}$ will be more concise. Precisely, define a band limited version of w, written as w_h, by

$$\hat{w}_h(m) = \begin{cases} \hat{w}(m), & \text{if } m\in B_{q-1}^d \\ 0, & \text{if } m\in (B_{q-1}^d)^c\,. \end{cases} \tag{5.7}$$

To estimate $\inf_{v\in\mathcal{F}_{t,q-1}} \|w-v\|_t^2$, we follow the two steps below:

Step 1: we prove $\|w - w_h\|_t^2 \lesssim 2^{-q(2t-d)}|y|^2$. Let us calculate the quantity directly:

$$\begin{aligned}
\|w - w_h\|_t^2 &= (4\pi^2)^{-t} \sum_{m\in(B_{q-1}^d)^c} |m|^{-2t}|g(m)|^2 \\
&= (4\pi^2)^{-t}\left(\sum_{m\in\mathbb{Z}^d\setminus\{0\}} |m|^{-2t}|g(m)|^2 - \sum_{m\in B_{q-1}^d} |m|^{-2t}|g(m)|^2\right) \\
&= (4\pi^2)^{-t}\left(\sum_{m\in B_q^d} M_q^t(m)|g(m)|^2 - \sum_{m\in B_{q-1}^d} |m|^{-2t}|g(m)|^2\right) \\
&\lesssim 2^{-2qt}\sum_{m\in B_q^d} |g(m)|^2 \lesssim 2^{-q(2t-d)}|y|^2.
\end{aligned}$$

Here we have used the fact that $M_q^t(m) - |m|^{-2t} \lesssim 2^{-2qt}$ for $m \in B_{q-1}^d$ and $M_q^t(m) \lesssim 2^{-2qt}$ for $m \in B_q^d\backslash B_{q-1}^d$, according to the results in Lemma 2.7. In the last line, the bound (5.6) is applied.

Step 2: We prove $\inf_{v\in\mathcal{F}_{t,q-1}} \|w_h - v\|_t^2 \lesssim 2^{-q(2t-d)}|y|^2$. Based on Theorem 2.5, we know the optimal v for this variational problem has the Fourier coefficients

$$\hat{v}(m) = \begin{cases} 0, & \text{if } m = 0 \\ |m|^{-2t}\frac{(T_{q-1}\hat{w}_h)(m)}{M_{q-1}^t(m)}, & \text{else}\,. \end{cases}$$

Then, using the Fourier representation of the norm, we get

$$\begin{aligned}
&\|w_h - v\|_t^2 \\
&\simeq \sum_{m\in\mathbb{Z}^d\setminus\{0\}} |m|^{2t}|\hat{w}_h(m) - \hat{v}(m)|^2 \\
&= \sum_{m\in B_{q-1}^d\setminus\{0\}} |m|^{-2t}|g(m) - \frac{(T_{q-1}\hat{w}_h)(m)}{M_{q-1}^t(m)}|^2 + \sum_{m\in(B_{q-1}^d)^c} |m|^{-2t}|\frac{(T_{q-1}\hat{w}_h)(m)}{M_{q-1}^t(m)}|^2\,.
\end{aligned}$$

For the first term, since w_h is band-limited, we know if $m \in B_{q-1}^d\backslash\{0\}$, then $(T_{q-1}\hat{w}_h)(m) = |m|^{-2t}g(m)$. Thus, we can write this term as

$$\begin{aligned}
&\sum_{m\in B_{q-1}^d\setminus\{0\}} |m|^{-2t}|g(m)|^2\Big(1 - \frac{|m|^{-2t}}{M_{q-1}^t(m)}\Big)^2 \\
&= \sum_{m\in B_{q-1}^d\setminus\{0\}} |m|^{-2t}|g(m)|^2\Big(\frac{M_{q-1}^t(m) - |m|^{-2t}}{M_{q-1}^t(m)}\Big)^2 \\
&\overset{a)}{\lesssim} \sum_{m\in B_{q-1}^d\setminus\{0\}} |m|^{-2t}|g(m)|^2\Big(\frac{2^{-4tq}}{|m|^{-4t}}\Big) \\
&\overset{b)}{\lesssim} \sum_{m\in B_{q-1}^d\setminus\{0\}} 2^{-2tq}|g(m)|^2 \lesssim 2^{-q(2t-d)}|y|^2
\end{aligned}$$

where in a), we have used the fact that $M_{q-1}^t(m) - |m|^{-2t} \simeq 2^{-2tq}$ and $M_{q-1}^t(m) \simeq |m|^{-2t}$ for $m \in B_{q-1}^d\backslash\{0\}$ based on Lemma 2.7. In b), we have used $|m| \lesssim 2^q$. The

last inequality is obtained by recalling (5.6).
For the second term, we write

$$\begin{aligned}
&\sum_{m\in(B^d_{q-1})^c} |m|^{-2t}|\frac{(T_{q-1}\hat{w}_h)(m)}{M^t_{q-1}(m)}|^2 \\
&= \sum_{m\in\mathbb{Z}^d\setminus\{0\}} |m|^{-2t}|\frac{(T_{q-1}\hat{w}_h)(m)}{M^t_{q-1}(m)}|^2 - \sum_{m\in B^d_{q-1}\setminus\{0\}} |m|^{-2t}|\frac{(T_{q-1}\hat{w}_h)(m)}{M^t_{q-1}(m)}|^2 \\
&\overset{c)}{=} \sum_{m\in B^d_{q-1}\setminus\{0\}} (M^t_{q-1}(m)-|m|^{-2t})|\frac{(T_{q-1}\hat{w}_h)(m)}{M^t_{q-1}(m)}|^2 \\
&= \sum_{m\in B^d_{q-1}\setminus\{0\}} (M^t_{q-1}(m)-|m|^{-2t})|\frac{|m|^{-2t}g(m)}{M^t_{q-1}(m)}|^2 \\
&\lesssim 2^{-2tq}\sum_{m\in B^d_{q-1}\setminus\{0\}} |g(m)|^2 \lesssim 2^{-q(2t-d)}|y|^2\,,
\end{aligned}$$

where in c), we have used the periodicity of the function $\frac{(T_{q-1}\hat{w}_h)(m)}{M^t_{q-1}(m)}$.

Now, combining Step 1 and 2 leads to the conclusion

$$\inf_{v\in\mathcal{F}_{t,q-1}} \|w-v\|_t^2 \lesssim 2^{-q(2t-d)}|y|^2\,,$$

and in particular, it implies

$$\lambda_{\max}(K(t,q)/K(t,q-1)) \lesssim 2^{-q(2t-d)}\,.$$

As a consequence of the upper and lower bounds for the eigenvalues of the matrix $K(t,q)/K(t,q-1)$, we deduce that they are all on the scale of $2^{-q(2t-d)}$. Let C be a constant independent of t,q such that $C^{-1}2^{-q(2t-d)} \preceq K(t,q)/K(t,q-1) \preceq C2^{-q(2t-d)}$. Then,

$$\begin{aligned}
(2^{qd}-2^{(q-1)d})((2t-d)(-q)\log 2 - C) &\le \log\det K(t,q)/K(t,q-1) \\
&\le (2^{qd}-2^{(q-1)d})((2t-d)(-q)\log 2 + C)\,.
\end{aligned}$$

Using the implied bounds on the recursion relation, we get

$$(2t-d)g_1(q) - Cg_2(q) + K(t,0) \le \log\det K(t,q) \le (2t-d)g_1(q) + Cg_2(q) + K(t,0)\,,$$

where $g_1(q) = \sum_{k=1}^q (2^{kd}-2^{(k-1)d})(-k\log 2)$ and $g_2(q) = (2^{qd}-1)(2t-d)$. Summing the series in $g_1(q)$ leads to $g_1(q) \simeq -q2^{qd}$. □

5.6. **Proof of Theorem 2.11.**

Proof. Recall the definition,

$$s^{\mathrm{EB}}(q) = \operatorname*{argmin}_t \mathsf{L}^{\mathrm{EB}}(t,q) := \|u(\cdot,t,q)\|_t^2 + \log\det K(t,q)\,.$$

Define a rescaled version of the loss function by

$$\tilde{L}_{\mathrm{EB}}(t,q) = \frac{1}{|g_1(q)|}\mathsf{L}^{\mathrm{EB}}(t,q) = \underbrace{\frac{1}{|g_1(q)|}\|u(\cdot,t,q)\|_t^2}_{①} + \underbrace{\frac{1}{|g_1(q)|}\log\det K(t,q)}_{②}\,.$$

We note that by Proposition 2.10, we have $|g_1(q)| \sim q2^{qd}$. Now, we estimate the growth rate of ① and ② separately. From Proposition 2.8 and 2.10, we get

$$① \simeq \underbrace{\frac{1}{q}2^{-q(2s-2t+d)}\xi_0^2}_{③} + \underbrace{\frac{1}{q}2^{-q(2s-2t)}\sum_{m\in B_q^d\setminus\{0\}} 2^{-q(2t-2s+d)}|m|^{2t-2s}\xi_m^2}_{④},$$

and for the log det part, it holds that

$$d-2t+\frac{-Cg_2(q)+K(t,0)}{|g_1(q)|} \le ② \le d-2t+\frac{Cg_2(q)+K(t,0)}{|g_1(q)|}.$$

It follows that $\lim_{q\to\infty} ② = d-2t$. Thus, our remaining task is to analyze terms ③, ④ in ①. We split the problem into four cases.

Case 1: $t=s$. It is easy to see $\lim_{q\to\infty} ③ = 0$ and

$$④ = \frac{1}{q}2^{-qd}\sum_{m\in B_q^d\setminus\{0\}}\xi_m^2 = \frac{1}{q}\alpha(d,q),$$

so that $\lim_{q\to\infty} ④ = 0$. Here we use the definition of α in Lemma 2.9. Therefore, $\lim_{q\to\infty}\tilde{L}_{\mathrm{EB}}(s,q) = d-2s$.

Case 2: $1/\delta \ge t \ge s+\epsilon$. We have ③ ≥ 0. The term ④ can be written as

$$④ = \frac{1}{q2^{-q(2t-2s)}}\alpha(2t-2s+d,q),$$

where we recall the definition of the function α in Lemma 2.9. According to this lemma, we get the uniform convergence

$$\lim_{q\to\infty}\alpha(2t-2s+d,q) = \gamma(2t-2s+d) > 0$$

in probability. In the meantime, $\lim_{q\to\infty} q2^{-q(2t-2s)} = 0$. So, $\lim_{q\to\infty} ④ = \infty$ in probability, and uniformly in $1/\delta \ge t \ge s+\epsilon$. In terms of $\tilde{L}_{\mathrm{EB}}(t,q)$, this corresponds to $\lim_{q\to\infty}\tilde{L}_{\mathrm{EB}}(t,q) = \infty$.

Case 3: $s-\epsilon \ge t \ge s-d/2+\epsilon$. In this case, $2t-2s+d \ge \epsilon$ so Lemma 2.9 can be applied. We write the term

$$④ = \frac{2^{-q(2s-2t)}}{q}\alpha(2t-2s+d,q).$$

This will converge to 0 as q goes to infinity, since $\lim_{q\to\infty}\frac{2^{-q(2s-2t)}}{q} = 0$ and $\lim_{q\to\infty}\alpha(2t-2s+d,q) = \gamma(2t-2s+d) \in (0,\infty)$. The term ③ also converges to 0. Thus, $\lim_{q\to\infty}\tilde{L}_{\mathrm{EB}}(t,q) = d-2t$ in probability, and uniformly for $s-\epsilon \ge t \ge s-d/2+\epsilon$.

Case 4: $s-d/2+\epsilon \ge t \ge d/2+\delta$. We still have that ③ converges to 0. For term ④, we have

$$④ = \frac{2^{-qd}}{q}\sum_{m\in B_q^d\setminus\{0\}}|m|^{2t-2s}\xi_m^2 \le \frac{2^{-qd}}{q}\sum_{m\in B_q^d\setminus\{0\}}|m|^{2(s-d/2+\epsilon)-2s}\xi_m^2$$

where we have used the monotonicity of the function $|m|^{2t-2s}$ with respect to t. Then, it reduces to the case $t = s-d/2+\delta$, which is covered by Case 3. Hence,

we have $\lim_{q\to\infty} ④ = 0$ uniformly for $s - d/2 + \delta \geq t \geq d/2 + \delta$. Therefore, we get $\lim_{q\to\infty} \tilde{L}_{\mathrm{EB}}(t,q) = d-2t$ in probability, and uniformly for $s-d/2+\delta \geq t \geq d/2+\delta$.

Let us make a summary of the arguments above. We have established that, for any small $\epsilon > 0$, $\lim_{q\to\infty} \tilde{L}_{\mathrm{EB}}(t,q) = \infty$ uniformly for $1/\delta \geq t \geq s+\epsilon$, and $\lim_{q\to\infty} \tilde{L}_{\mathrm{EB}}(t,q) = d-2t$ uniformly for $s-\epsilon \geq t \geq d/2+\delta$, and $\lim_{q\to\infty} \tilde{L}_{\mathrm{EB}}(s,q) = d-2s$. All the convergence is in probability. Note that s^{EB} is the minimizer of $L_{\mathrm{EB}}(t,q)$, hence also of $\tilde{L}_{\mathrm{EB}}(t,q)$. The above convergence results for $\tilde{L}_{\mathrm{EB}}(t,q)$ imply that $s^{\mathrm{EM}} \in (s-\epsilon, s+\epsilon)$ with probability 1 as q goes to infinity, for any $\epsilon > 0$. Thus, we must have

$$\lim_{q\to\infty} s^{\mathrm{EB}}(q) = s\,.$$

The proof is complete. □

5.7. **Proof of Proposition 2.13.**

Proof. In order to write the interaction terms as a random series with some desired independence pattern for the random variables involved, we need to consider the geometry of the lattice carefully. We introduce another set $S_q := \{m \in \mathbb{Z} : -2^{q-2} \leq m \leq 3\times 2^{q-2} - 1\}$ and let $S_q^d = S_q \otimes S_q \otimes \cdots \otimes S_q$ denote the tensor product of d multiples of S_q. The set S_q is a shift of B_q, and S_q^d is a shift of B_q^d.

Define the set $B_{q-1}^d + 2^{q-1}k := \{m + 2^{q-1}k : m \in B_{q-1}^d\}$ for $k \in \mathbb{Z}^d$. We have the relation

$$S_q^d = \bigcup_{k\in\mathbb{Z}_2^d} (B_{q-1}^d + 2^{q-1}k)$$

where $\mathbb{Z}_2^d = \{0,1\}^d$. Note that for $k_1 \neq k_2$, the intersection between $B_{q-1}^d + 2^{q-1}k_1$ and $B_{q-1}^d + 2^{q-1}k_2$ is empty.

Using (2.8) and the periodicity of the functions involved, we get

$$\begin{aligned}
&\|u(\cdot,t,q) - u(\cdot,t,q-1)\|_t^2 \\
=&(4\pi^2)^t \sum_{m\in B_q^d} M_q^t(m) \left(\frac{T_q\hat{u}(m)}{M_q^t(m)} - \frac{T_{q-1}\hat{u}(m)}{M_{q-1}^t(m)}\right)^2 \\
=&(4\pi^2)^t \sum_{m\in S_q^d} M_q^t(m) \left(\frac{T_q\hat{u}(m)}{M_q^t(m)} - \frac{T_{q-1}\hat{u}(m)}{M_{q-1}^t(m)}\right)^2 \\
=&(4\pi^2)^t \sum_{k\in\mathbb{Z}_2^d} \sum_{m\in(B_{q-1}^d + 2^{q-1}k)} M_q^t(m) \left(\frac{T_q\hat{u}(m)}{M_q^t(m)} - \frac{T_{q-1}\hat{u}(m)}{M_{q-1}^t(m)}\right)^2 .
\end{aligned}$$

Recall the relation

$$T_{q-1}\hat{u}(m) = \sum_{l\in\mathbb{Z}_2^d} T_q\hat{u}(m + 2^{q-1}l)\,,$$

based on which we get

$$\begin{aligned}
&\frac{T_q\hat{u}(m)}{M_q^t(m)} - \frac{T_{q-1}\hat{u}(m)}{M_{q-1}^t(m)} \\
=&(\frac{1}{M_q^t(m)} - \frac{1}{M_{q-1}^t(m)})T_q\hat{u}(m) - \frac{1}{M_{q-1}^t} \sum_{l\in\mathbb{Z}_2^d\setminus\{0\}} T_q\hat{u}(m + 2^{q-1}l)\,.
\end{aligned}$$

Since $u^\dagger \sim \mathcal{N}(0,(-\Delta)^{-s})$, it holds $\hat{u}(m) \sim \mathcal{N}(0,(4\pi^2)^{-s}|m|^{-2s})$. Moreover, for different m, these Gaussian random variables are independent from each other. Thus, for a fixed k and for $m \in (B^d_{q-1} + 2^{q-1}k)$, the Gaussian random variables

$$\frac{T_q\hat{u}(m)}{M^t_q(m)} - \frac{T_{q-1}\hat{u}(m)}{M^t_{q-1}(m)}$$

are independent from each other. Furthermore, by calculating their variance, we can write

$$\begin{aligned}
&M^t_q(m)\left(\frac{T_q\hat{u}(m)}{M^t_q(m)} - \frac{T_{q-1}\hat{u}(m)}{M^t_{q-1}(m)}\right)^2 \\
=&(4\pi^2)^{-s}\left[(\frac{1}{M^t_q(m)} - \frac{1}{M^t_{q-1}(m)})^2 M^t_q(m)M^s_q(m) + \frac{M^t_q(m)}{(M^t_{q-1}(m))^2}\sum_{l\in\mathbb{Z}^d_2\setminus\{0\}} M^s_q(m+2^{q-1}l)\right]\xi^2_{k,m} \\
=&(4\pi^2)^{-s}\left[\frac{M^s_q(m)(M^t_q(m)-M^t_{q-1}(m))^2}{M^t_q(m)(M^t_{q-1}(m))^2} + \frac{M^t_q(m)}{(M^t_{q-1}(m))^2}\sum_{l\in\mathbb{Z}^d_2\setminus\{0\}} M^s_q(m+2^{q-1}l)\right]\xi^2_{k,m} \\
=:& A_{k,m}\xi^2_{k,m}
\end{aligned}$$

where $\{\xi_{k,m}\}_m$ are independent unit scalar Gaussian random variables. Clearly, we have the lower bound

$$A_{k,m} \geq (4\pi^2)^{-s}\frac{M^t_q(m)}{(M^t_{q-1}(m))^2}M^s_q(m-2^{q-1}k)\,.$$

Thus, denoting $e_1 = (1,0,...,0) \in \mathbb{Z}^d$, we get

$$\begin{aligned}
&\|u(\cdot,t,q)-u(\cdot,t,q-1)\|^2_t \\
\geq&(4\pi^2)^{t-s}\sum_{k\in\mathbb{Z}^d_2}\sum_{m\in(B^d_{q-1}+2^{q-1}k)}\frac{M^t_q(m)}{(M^t_{q-1}(m))^2}M^s_q(m-2^{q-1}k)\xi^2_{k,m} \\
\geq&(4\pi^2)^{t-s}\sum_{m\in(B^d_{q-1}+2^{q-1}e_1)}\frac{M^t_q(m)}{(M^t_{q-1}(m))^2}M^s_q(m-2^{q-1}e_1)\xi^2_{e_1,m} \\
=&(4\pi^2)^{t-s}\sum_{m\in(B^d_{q-1}+2^{q-1}e_1)}\frac{M^t_q(m)}{(M^t_{q-1}(m-2^{q-1}e_1))^2}M^s_q(m-2^{q-1}e_1)\xi^2_{e_1,m} \\
\gtrsim&\sum_{m\in(B^d_{q-1}\setminus\{0\}+2^{q-1}e_1)}\frac{2^{-2qt}}{|m-2^{q-1}e_1|^{-4t}}|m-2^{q-1}e_1|^{2s}\xi^2_{e_1,m} \\
=&\sum_{m\in B^d_{q-1}\setminus\{0\}} 2^{-2qt}|m|^{4t-2s}\xi^2_{e_1,m+2^{q-1}e_1}\,.
\end{aligned}$$

In the above derivation, we have used the fact that for $m \in B^d_{q-1}$, it holds that $M^s_q(m) \simeq |m|^{-2s}, M^t_{q-1}(m) \simeq |m|^{-2t}$, and in particular, $M^t_q(m) \simeq |m|^{-2t} \simeq 2^{-2qt}$ for $m \in (B^d_{q-1}\setminus\{0\} + 2^{q-1}e_1)$. Renaming the subscripts in $\xi_{e_1,m+2^{q-1}e_1}$ completes the proof. □

5.8. **Proof of Proposition 2.14.**

Proof. We need to upper bound $A_{k,m}$ for $k \in \mathbb{Z}_2^d, m \in B_{q-1}^d + 2^{q-1}k$, which is defined in the proof of Proposition 2.13. First, we have

$$\sum_{l\in\mathbb{Z}_2^d\setminus\{0\}} M_q^s(m+2^{q-1}l) = M_{q-1}^s(m) - M_q^s(m)\,,$$

and the estimate $0 \le M_{q-1}^t(m) - M_q^t(m) \le M_{q-1}^t(m)$ for any $d/2+\delta \le t \le 1/\delta$. Based on this observation, for $k \in \mathbb{Z}^d\setminus\{0\}$ and $m \in B_{q-1}^d\setminus\{0\} + 2^{q-1}k$, we have the bound

$$\begin{aligned} A_{k,m} &\lesssim \frac{M_q^s(m)}{M_q^t(m)} + M_q^t(m)\frac{M_{q-1}^s(m)}{(M_{q-1}^t(m))^2} \\ &\lesssim 2^{-q(2s-2t)} + 2^{-2tq}|m-2^{q-1}k|^{4t-2s} \end{aligned}$$

where we have used the fact that for $m \in B_{q-1}^d\setminus\{0\}+2^{q-1}k$, it holds that $M_q^s(m) \simeq 2^{-2sq}, M_q^t(m) \simeq 2^{-2tq}, M_{q-1}^s(m) \simeq |m-2^{q-1}k|^{-2s}, M_{q-1}^t(m) \simeq |m-2^{q-1}k|^{-2t}$, according to Lemma 2.7. For $m = 2^{q-1}k$, we get $A_{k,m} \lesssim 2^{-q(2s-2t)}$. So in general, we can write $A_{k,m} \lesssim 2^{-q(2s-2t)} + 2^{-2tq}|m-2^{q-1}k|^{4t-2s}$ for $m \in B_{q-1}^d + 2^{q-1}k$ where we use the convention that $|m|^\alpha = 0$ for $m = 0$ and any $\alpha \in \mathbb{R}$ to make the notation more compact.

When $k = 0$, using Lemma 2.7 again, we get for $m \in B_{q-1}^d\setminus\{0\}$,

$$\begin{aligned} A_{k,m} &\lesssim \frac{M_q^s(m)(M_q^t(m)-M_{q-1}^t(m))^2}{M_q^t(m)(M_{q-1}^t(m))^2} + \frac{M_q^t(m)}{(M_{q-1}^t(m))^2}(M_{q-1}^s(m)-M_q^s(m)) \\ &\lesssim \frac{|m|^{-2s}2^{-4tq}}{|m|^{-6t}} + \frac{|m|^{-2t}}{|m|^{-4t}}2^{-2sq} \\ &= |m|^{6t-2s}2^{-4tq} + |m|^{2t}2^{-2sq} \\ &\lesssim 2^{-2tq}|m|^{4t-2s} + 2^{-q(2s-2t)}\,, \end{aligned}$$

where in the last line we used the relation $|m| \lesssim 2^q$. For $m = 0$, based on the above calculation, we can get $A_{k,m} \lesssim 2^{-q(2s-2t)}$. Thus, generally, we can write $A_{k,m} \lesssim 2^{-2tq}|m|^{4t-2s} + 2^{-q(2s-2t)}$ for $m \in B_{q-1}^d$ by using the notational convention above.

Combining these estimates, we arrive at

$$\begin{aligned} &\|u(\cdot,t,q)-u(\cdot,t,q-1)\|_t^2 \\ &\lesssim \sum_{k\in\mathbb{Z}_2^d}\sum_{m\in(B_{q-1}^d+2^{q-1}k)} A_{k,m}\xi_{k,m}^2 \\ &\lesssim \sum_{k\in\mathbb{Z}_2^d}\sum_{m\in(B_{q-1}^d+2^{q-1}k)} (2^{-q(2s-2t)} + 2^{-2tq}|m-2^{q-1}k|^{4t-2s})\xi_{k,m}^2 \\ &= \sum_{k\in\mathbb{Z}_2^d}\sum_{m\in B_{q-1}^d} (2^{-q(2s-2t)} + 2^{-2tq}|m|^{4t-2s})\xi_{k,m+2^{q-1}k}^2\,. \end{aligned}$$

After a change of notation, we get the desired estimate. □

5.9. **Proof of Theorem 2.15.**

Proof. Recall

$$s^{\mathrm{KF}}(q) = \operatorname*{argmin}_{t\in[d/2+\delta,1/\delta]} \mathsf{L}^{\mathrm{KF}}(t,q) := \frac{\|u(\cdot,t,q)-u(\cdot,t,q-1)\|_t^2}{\|u(\cdot,t,q)\|_t^2}\,.$$

We analyze the denominator and numerator separately. We start with the numerator. Let

$$V_1(t,q) = \frac{1}{q}2^{q(s-d/2)}\|u(\cdot,t,q)-u(\cdot,t,q-1)\|_t^2\,.$$

Case 1: $t=\frac{s-d/2}{2}$. We derive an upper bound on V_1. By Proposition 2.14,

$$\|u(\cdot,t,q)-u(\cdot,t,q-1)\|_t^2 \lesssim \sum_{k\in\mathbb{Z}_2^d}\sum_{m\in B_{q-1}^d}(2^{-q(2s-2t)}+2^{-2tq}|m|^{4t-2s})\xi_{k,m}^2\,.$$

Take $t=\frac{s-d/2}{2}$. For each $k\in\mathbb{Z}_2^d$, consider the term

$$\begin{aligned}
V_1^k(t,q) &= \frac{1}{q}2^{q(s-d/2)}\sum_{m\in B_{q-1}^d}(2^{-q(2s-2t)}+2^{-2tq}|m|^{4t-2s})\xi_{k,m}^2\\
&= \frac{1}{q}2^{q(s-d/2)}\sum_{m\in B_{q-1}^d}(2^{-q(s+d/2)}+2^{-q(s-d/2)}|m|^{-d})\xi_{k,m}^2\\
&= \frac{1}{q}\sum_{m\in B_{q-1}^d}(2^{-qd}+|m|^{-d})\xi_{k,m}^2\\
&\lesssim \frac{1}{q}\sum_{m\in B_{q-1}^d}|m|^{-d}\xi_{k,m}^2\,.
\end{aligned}$$

By Lemma 2.9, $\lim_{q\to\infty}\frac{1}{q}\sum_{m\in B_{q-1}^d}|m|^{-d}\xi_{k,m}^2=\gamma(0)\in(0,\infty)$. Thus, $V_1^k(t,q)$ remains bounded for $q\in\mathbb{N}$. Since $V_1(t,q)=\sum_{k\in\mathbb{Z}_2^d}V_1^k(t,q)$, it follows that $V_1(t,q)$ remains bounded for $q\in\mathbb{N}$, in the case $t=\frac{s-d/2}{2}$.

Case 2: $1/\delta\geq t\geq\frac{s-d/2}{2}+\epsilon$. We provide a lower bound of V_1 here. Using Proposition 2.13, we get

$$\begin{aligned}
V_1(t,q) &\gtrsim \frac{1}{q}2^{q(s-d/2)}\sum_{m\in B_{q-1}^d\setminus\{0\}}2^{-2tq}|m|^{4t-2s}\xi_m^2\\
&= \frac{1}{q}2^{q(s-d/2-2t)}\sum_{m\in B_{q-1}^d\setminus\{0\}}|m|^{4t-2s}\xi_m^2\\
&= \frac{1}{q}2^{q(s-d/2-2t)}2^{(q-1)(4t-2s+d)}\cdot\left(2^{-(q-1)(4t-2s+d)}\sum_{m\in B_{q-1}^d\setminus\{0\}}|m|^{4t-2s}\xi_m^2\right)\\
&= \frac{1}{q}2^{(q/2-1)(4t-2s+d)}\alpha(4t-2s+d,q-1)\,.
\end{aligned}$$

By Lemma 2.9, $\lim_{q\to\infty}\alpha(4t-2s+d,q-1)=\gamma(4t-2s+d)>0$ uniformly for $1/\delta\geq t\geq\frac{s-d/2}{2}+\epsilon$. Since $\lim_{q\to\infty}\frac{1}{q}2^{(q/2-1)(4t-2s+d)}=\infty$, we get $\lim_{q\to\infty}V_1(t,q)=\infty$ and its growth rate is $\gtrsim\frac{1}{q}2^{(q/2-1)(4t-2s+d)}$.

Case 3: $\frac{s-d/2}{2} - \epsilon \geq t \geq d/2 + \delta$. We provide a lower bound on V_1 here. Similarly to our analysis in Case 2, we have

$$\begin{aligned} V_1(t,q) &\gtrsim \frac{1}{q} 2^{q(s-d/2-2t)} \sum_{m \in B^d_{q-1} \setminus \{0\}} |m|^{4t-2s} \xi_m^2 \\ &\gtrsim \frac{1}{q} 2^{q(s-d/2-2t)} \xi_1^2 \,. \end{aligned}$$

Then, it holds that

$$\mathbb{P}(\frac{1}{q} 2^{q(s-d/2-2t)} \xi_1^2 \geq 2^{q(s-d/2-2t)/2}) = \mathbb{P}(\xi_1^2 \geq q 2^{-q(s-d/2-2t)/2}) \to 1$$

as $q \to \infty$. Thus, we get $\lim_{q\to\infty} V_1(t,q) = \infty$ uniformly for this range of t and the growth rate is $\gtrsim 2^{q(s-d/2-2t)/2}$. We have finished the analysis of the numerator. Now we proceed to analyze the denominator, which comprises the norm term. From Proposition 2.8, we have

$$\|u(\cdot,t,q)\|_t^2 \simeq 2^{-q(2s-2t)} \xi_0^2 + \sum_{m \in B^d_q \setminus \{0\}} |m|^{2t-2s} \xi_m^2 \,, \tag{5.8}$$

where $\{\xi_m\}_{m \in B^d_q}$ are independent unit scalar Gaussian random variables. Recall that our final target in this theorem is to show that, for any $\epsilon > 0$,

$$\lim_{q\to\infty} \mathbb{P}[s^{\mathrm{KF}}(q) \in (\frac{s-d/2}{2} - \epsilon, \frac{s-d/2}{2} + \epsilon)] = 1 \,.$$

Let $I_\epsilon = [d/2+\delta, 1/\delta]/[\frac{s-d/2}{2} - \epsilon, \frac{s-d/2}{2} + \epsilon]$. By rewriting the loss function, it suffices to show

$$\lim_{q\to\infty} \mathbb{P}[\frac{V_1(\frac{s-d/2}{2}, q)}{\|u(\cdot, \frac{s-d/2}{2}, q)\|^2_{\frac{s-d/2}{2}}} \geq \inf_{t \in I_\epsilon} \frac{V_1(t,q)}{\|u(\cdot,t,q)\|_t^2}] = 0 \,.$$

Let us write

$$r(t,q) = \frac{V_1(t,q)}{V_1(\frac{s-d/2}{2}, q)} \cdot \frac{\|u(\cdot, \frac{s-d/2}{2}, q)\|^2_{\frac{s-d/2}{2}}}{\|u(\cdot,t,q)\|_t^2} \,, \tag{5.9}$$

then all we need is to show

$$\lim_{q\to\infty} \mathbb{P}[\inf_{t \in I_\epsilon} r(t,q) \leq 1] = 0 \,.$$

For $t \in I^1_\epsilon = [d/2+\delta, \frac{s-d/2}{2} - \epsilon]$, according to the analysis for the numerator, we have that for some constant C independent of q,

$$\lim_{q\to\infty} \mathbb{P}[\inf_{t \in I^1_\epsilon} \frac{V_1(t,q)}{2^{q(s-d/2-2t)/2}} \geq C] = 1 \,, \tag{5.10}$$

and also, $V_1(\frac{s-d/2}{2}, q)$ remains uniformly bounded for $q \in \mathbb{N}$. Furthermore, the equation (5.8) implies the following relation:

$$\inf_{t \in I^1_\epsilon} \frac{\|u(\cdot, \frac{s-d/2}{2}, q)\|^2_{\frac{s-d/2}{2}}}{\|u(\cdot,t,q)\|_t^2} \gtrsim 1 \,, \tag{5.11}$$

due to the inequality $t \leq \frac{s-d/2}{2} - \epsilon$. Combining the above two estimates in (5.10)(5.11), and recalling the expression for $r(t,q)$ in (5.9), we get

$$\lim_{q\to\infty} \mathbb{P}[\inf_{t\in I_\epsilon^1} r(t,q) \leq 1] = 0\,. \tag{5.12}$$

Then, let $I_\epsilon^2 = [\frac{s-d/2}{2} + \epsilon, 1/\delta]$. We also need to show $\lim_{q\to\infty} \mathbb{P}[\inf_{t\in I_\epsilon^2} r(t,q) \leq 1] = 0$, or equivalently,

$$\lim_{q\to\infty} \mathbb{P}[\frac{\|u(\cdot, \frac{s-d/2}{2}, q)\|^2_{\frac{s-d/2}{2}}}{V_1(\frac{s-d/2}{2}, q)} \leq \sup_{t\in I_\epsilon^2} \frac{\|u(\cdot,t,q)\|_t^2}{V_1(t,q)}] = 0\,.$$

Since $V_1(\frac{s-d/2}{2}, q)$ remains bounded according to the result in the above Case 1, it suffices to show

$$\lim_{q\to\infty} \sup_{t\in I_\epsilon^2} \frac{\|u(\cdot,t,q)\|_t^2}{V_1(t,q)} = 0$$

in probability. Using the estimate of $V_1(t,q)$ in Case 2 that $V_1(t,q) \gtrsim \frac{1}{q} 2^{(q/2-1)(4t-2s+d)}$, it suffices to show

$$\lim_{q\to\infty} \sup_{t\in I_\epsilon^2} q 2^{-(q/2-1)(4t-2s+d)} \|u(\cdot,t,q)\|_t^2 = 0\,.$$

To achieve this, we recall the expression of the norm term and write

$$\begin{aligned} &q2^{-(q/2-1)(4t-2s+d)}\|u(\cdot,t,q)\|_t^2 \\ \simeq &q2^{-q(s+d)+4t-2s+d}\xi_0^2 + q2^{-(q/2-1)(4t-2s+d)} \sum_{m\in B_q^d\setminus\{0\}} |m|^{2t-2s}\xi_m^2 \end{aligned}$$

Clearly, the first term on the right hand side converges to 0, so we only need to deal with the second term. Let

$$\beta(t,q) = q2^{-(q/2-1)(4t-2s+d)} \sum_{m\in B_q^d\setminus\{0\}} |m|^{2t-2s}\xi_m^2\,.$$

Consider $t \in [s - d/2 + \epsilon', 1/\delta]$ where ϵ' is a parameter to be tuned. We have $2t - 2s + d \geq \epsilon' > 0$ so we are able to write

$$\begin{aligned} \beta(t,q) &= q2^{-(q/2-1)(4t-2s+d)} 2^{q(2t-2s+d)} \alpha(2t-2s+d, q) \\ &= q2^{-q(s-d/2)+4t-2s+d} \alpha(2t-2s+d, q)\,. \end{aligned}$$

By Lemma 2.9, $\lim_{q\to\infty} \alpha(2t-2s+d,q) = \gamma(2t-2s+d)$ in probability uniformly for $t \in [s-d/2+\epsilon', 1/\delta]$. Since $\lim_{q\to\infty} q2^{-(q/2-1)(4t-2s+d)}2^{q(2t-2s+d)} = 0$, we get $\lim_{q\to\infty} \sup_{t\in[s-d/2+\epsilon',1/\delta]} \beta(t,q) = 0$.
For $t \in [\frac{s-d/2}{2} + \epsilon, s - d/2 + \epsilon']$, we have the estimate

$$q2^{-(q/2-1)(4t-2s+d)} \leq \left(q2^{-(q/2-1)(4t-2s+d)}\right)_{t=\frac{s-d/2}{2}+\epsilon} = q2^{-2q\epsilon+4\epsilon}$$

and

$$\sum_{m\in B_q^d\setminus\{0\}} |m|^{2t-2s}\xi_m^2 \leq \sum_{m\in B_q^d\setminus\{0\}} |m|^{-d+2\epsilon'}\xi_m^2$$

where we have used the fact that t is upper bounded by $s - d/2 + \epsilon'$. Hence,

$$\sup_{t\in[\frac{s-d/2}{2}+\epsilon, s-d/2+\epsilon']} \beta(t,q) \le q2^{-2q\epsilon+4\epsilon} \sum_{m\in B_q^d\setminus\{0\}} |m|^{-d+2\epsilon'}\xi_m^2$$
$$= q2^{-2q\epsilon+4\epsilon}2^{2q\epsilon'}\alpha(2\epsilon', q)\,.$$

Now, we set $\epsilon' = \epsilon/2$ such that $\lim_{q\to\infty} q2^{-2q\epsilon+4\epsilon}2^{2q\epsilon'} = 0$. Lemma 2.9 leads to $\lim_{q\to\infty}\alpha(2\epsilon', q) = \gamma(2\epsilon') < \infty$, from which we can conclude $\lim_{q\to\infty}\sup_{t\in I_\epsilon^2}\beta(t,q) = 0$. Therefore, we get

$$\lim_{q\to\infty} \mathbb{P}[\inf_{t\in I_\epsilon^2} r(t,q) \le 1] = 0\,. \tag{5.13}$$

Combining (5.12) and (5.13) gives

$$\lim_{q\to\infty} \mathbb{P}[\inf_{t\in I_\epsilon} r(t,q) \le 1] = 0\,. \tag{5.14}$$

Based on the definition of $r(t,q)$ in (5.9) and the arguments therein, we obtain

$$\lim_{q\to\infty} \mathbb{P}[s^{\mathrm{KF}}(q) \in (\frac{s-d/2}{2} - \epsilon, \frac{s-d/2}{2} + \epsilon)] = 1\,,$$

from which the consistency of the KF estimator follows. □

Applied and Computational Mathematics, Caltech, 91106
E-mail address: yifanc@caltech.edu

Applied and Computational Mathematics, Caltech, 91106
E-mail address: owhadi@caltech.edu

Applied and Computational Mathematics, Caltech, 91106
E-mail address: astuart@caltech.edu