SLEIPNIR: Deterministic and Provably Accurate Feature Expansion for Gaussian Process Regression with Derivatives

Emmanouil Angelis* ANGELISE@ETHZ.CH
Learning and Adaptive Systems Group
ETH Zürich
Philippe Wenk* WENKPH@ETHZ.CH
Learning and Adaptive Systems Group
ETH Zürich and Max Planck ETH Center for Learning Systems
Bernhard Schölkopf BERNHARD.SCHOELKOPF@TUEBINGEN.MPG.DE
Empirical Inference Group
Max Planck Institute for Intelligent Systems, Tübingen
Stefan Bauer STEFAN.BAUER@TUEBINGEN.MPG.DE
Empirical Inference Group
Max Planck Institute for Intelligent Systems, Tübingen
Andreas Krause KRAUSEA@ETHZ.CH
Learning and Adaptive Systems Group
ETH Zürich

Abstract

Gaussian processes are an important regression tool with excellent analytic properties which allow for direct integration of derivative observations. However, vanilla GP methods scale cubically in the amount of observations. In this work, we propose a novel approach for scaling GP regression with derivatives based on quadrature Fourier features. We then prove deterministic, non-asymptotic and exponentially fast decaying error bounds which apply for both the approximated kernel as well as the approximated posterior. To furthermore illustrate the practical applicability of our method, we then apply it to ODIN, a recently developed algorithm for ODE parameter inference. In an extensive experiments section, all results are empirically validated, demonstrating the speed, accuracy, and practical applicability of this approach.

1. Introduction

Gaussian process (GP) regression (Rasmussen and Williams, 2006) is an important machine learning model with many desirable properties. Due to their inherently Bayesian nature, GPs naturally provide uncertainty estimates which are of crucial importance in Bayesian optimization (Mockus, 2012) or active learning (Seo et al., 2000). Furthermore, since a derivative of a GP is again a GP, there is a natural extension to derivative observations (Solak et al., 2003), which can be leveraged in the context of probabilistic numerics (Hennig et al., 2015) or Gaussian-process-based gradient matching (Calderhead et al., 2009; Wenk et al., 2020). Despite these appealing properties, a clear drawback of classic GP regression is the fact that they scale cubic in the amount of observation points.

∗. The first two authors contributed equally.

Scaling standard GP regression For standard, unconstrained GP regression, there exist several ideas on how to tackle this problem. One family of approaches focuses on summarizing the data set with a fixed amount of pseudo-observations, the so-called inducing points (Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006; Titsias, 2009). Complementary to these methods, special structure in either the kernel function (Wilson et al., 2014) or the input data (Cunningham et al., 2008) can be exploited to speed up the necessary matrix-vector multiplications needed to perform GP regression. Wilson and Nickisch (2015) combine these two ideas, creating an efficient approximation scheme linear in the amount of observations. Independently, Sarkka et al. (2013) propose a SDE-based reformulation and connect GP regression to Kalman filtering. Finally, there is a family of approaches approximating the kernel function via a finite dimensional scalar product of feature vectors. These features have been obtained using the Nyström method (Williams and Seeger, 2001), MC samples (Rahimi and Recht, 2008), sparse spectrum approximations (Lazaro-Gredilla et al., 2010), variational optimization (Hensman et al., 2017) or a quadrature scheme (Mutny and Krause, 2018). Furthermore, Solin and Särkkä develop a deterministic feature expansion based on the Fourier transform of the Laplace operator. While they are able to provide deterministic, non-asymptotic error bounds, their error decays linearly with the the size of the domain of the operator expansion L, which cannot grow faster than the amount of features. This essentially means that their approximation decays at best linearly.

Scaling GP regression with derivatives In the context of Gaussian process regression with derivatives though, scalable methods seem to have received little attention (Eriksson et al., 2018), despite their practical relevance. While some approaches like inducing points could be applied by just using the same inducing points, it is not obvious how such changes would affect the quality of the resulting estimates. Similarly, an empirical extension of the approach of Solin and Särkkä to the derivative case is presented by Solin et al. (2018), without quantifying the errors of the approximation scheme. For random Fourier features (RFF), Szabó and Sriperumbudur (2018) present a rigorous theoretical analysis. However, due to the probabilistic nature of RFF, it is not possible to provide deterministic, non-asymptotic guarantees for any given data set of fixed size.

To obtain deterministic (i.e. hold with probability 1) and non-asymptotic (i.e. for data sets with a fixed amount of observations) error bounds, we thus turn to quadratic Fourier features. Mutny and Krause (2018) recently derived exponentially fast decaying bounds for standard GP regression. Building on their work, we derive approximations and bounds for kernel derivatives. As we will demonstrate, these bounds can be used directly to quantify the absolute error of the predictive posterior mean and covariance of a GP with derivative information, leading to deterministic, non-asymptotic error bounds that can guide the choice of the amount of features needed to obtain a desired accuracy.

However, as we shall demonstrate, our bounds are so powerful that they can easily be extended to more applications. We will demonstrate this using ODE-informed regression (ODIN) as an example. In ODE-Informed regression, Gaussian processes are used to infer the parameters of a system of ODEs governing the dynamics of a time-continuous system that is observed at discrete time points under additive Gaussian noise. The idea to use GPs for this task goes back to the pioneering work of Calderhead et al. (2009), whose theoretical models were later refined by Dondelinger et al. (2013) and Wenk et al. (2018). As shown e.g. by Gorbach et al. (2017) and Abbati et al. (2019), such GP based inference schemes scale almost linearly in the dimension of the ODE. Since applications of these algorithms involve deducing novel scientific knowledge (Dony et al., 2019; Macdonald and

Husmeier, 2015), any approximation scheme would require a rigorous error analysis. Fortunately, our algorithm leads to deterministic, finite sample error bounds for the risk it optimizes, again guiding the choice of the complexity of the approximation for a desired accuracy level.

Contributions: In summary, we

- extend the QFF framework to Gaussian process regression with derivative information, naming the extension SLEIPNIR,
- derive, prove and empirically validate deterministic, finite-sample guarantees for the accuracy of said approximation,
- demonstrate how these theoretical insights can be used to control the approximation error in GPR with derivatives,
- expand these insights to consistently reduce the cubic run time of ODE informed regression (Wenk et al., 2020),
- verify the theoretical results empirically on four different systems with both locally linear and more involved, nonlinear dynamics, demonstrating a significant reduction in computational complexity without noticeable loss of accuracy.

All code needed to recreate our results can be found at `https://github.com/sdi1100041/SLEIPNIR` to facilitate future research and reproducibility.

2. Background

In this section, we provide a high-level overview of the background for our work. For an in-depth introduction, see Rasmussen and Williams (2006) (GPs), Rahimi and Recht (2008) (RFF) and Mutny and Krause (2018) (QFF).

2.1 Feature Expansions for normal GPR

Unfortunately, any reasoning in such a model requires the calculation and inversion of the covariance matrices $\boldsymbol{C}_\phi$ and $\boldsymbol{A}$. Without any tricks or approximations, this will scale as $\mathcal{O}(N^3)$. Random Fourier features (RFF) introduced by Rahimi and Recht (2008) and quadrature Fourier features (QFF) introduced by Mutny and Krause (2018) are two approximations that can be used to reduce this complexity for standard GP regression. Both approaches are based on the following observation: Any kernel represents a scalar product in a potentially infinite dimensional feature space, but if we allow for a small error, this could be approximated by a finite-dimensional feature vector.

For readability, we will introduce all concepts using a kernel with scalar inputs $k(t_i, t_j)$. However, it should be noted that this is by no means a necessary condition, and the concept generalizes nicely to higher dimensional inputs.

For a stationary, scalar kernel, Bochner's theorem (see e.g. Rudin, 1976) guarantees the existence of a density $p(\omega)$ such that we can write

$$k(|t_i - t_j|) = \int_{-\infty}^{\infty} p(\omega) \begin{pmatrix} \cos(\omega t_i) \\ \sin(\omega t_i) \end{pmatrix}^T \begin{pmatrix} \cos(\omega t_j) \\ \sin(\omega t_j) \end{pmatrix} d\omega. \tag{1}$$

This equation can be interpreted as a scalar product of infinitely many features given by $\cos(\omega t_i)$ and $\sin(\omega t_i)$. But how do we find the most important features such that the kernel k can be reasonably well approximated by a finite feature vector

$$k(|t_i - t_j|) \approx \boldsymbol{\phi}(t_i)^T \boldsymbol{\phi}(t_j)? \tag{2}$$

To obtain random Fourier features, Rahimi and Recht (2008) propose two different Monte Carlo sampling schemes. The first approximation just obtains MC samples from $p(\omega)$ and for each sample adds $[\sin(\omega t_i)\cos(\omega t_i)]$ to the feature vector $\boldsymbol{\phi}(t_i)$. In the second scheme, they observe that

$$\begin{pmatrix}\cos(\omega t_i)\\ \sin(\omega t_i)\end{pmatrix}^T \begin{pmatrix}\cos(\omega t_j)\\ \sin(\omega t_j)\end{pmatrix} = \mathbb{E}_b\{\sqrt{2}\cos(\omega t_i + b)\}, \tag{3}$$

where $b \sim \text{Unif}([0, 2\pi])$. Thus, one can obtain a second approximation by sampling concurrently b and ω and for each sample adding $\cos(\omega t_i + b)$ to the feature vector $\boldsymbol{\phi}(t_i)$. Accentuating the presence of the bias term b, we will refer to this approximation in the following as RFF-B, while we refer to the first one as RFF.

One main drawback of RFFs however is the fact that the intermediate sampling step makes it impossible to obtain any deterministic bounds on their approximation quality. Mutny and Krause (2018) thus propose to approximate the integral of Equation (1) using Hermitian quadrature. While RFF and RFF-B use sampling to determine the locations of the ω_i, in the context of Hermitian quadrature, these locations are fully determined by the functional form of the kernel. Thus, quadrature Fourier features (QFF) are deterministic in nature, leading to a deterministic, theoretical analysis. Furthermore, as a numerical integration scheme, they are especially efficient in low dimensions. As we shall see both in our theoretical and empirical analysis, similar properties can be achieved if we want to include derivative observations.

2.2 Gaussian Process Regression with Derivatives

Assume there exists a scalar-valued function $x(t)$. At N distinct time points $\boldsymbol{t} = [t_0, \ldots, t_{N-1}]$, we obtain noisy observations of the function itself and its derivatives, assuming Gaussian noise with standard deviations σ and $\sqrt{\gamma}$. Following Wenk et al. (2020), these observations are represented as the vectors $\boldsymbol{y}$ and $\boldsymbol{F}$. Using GP regression, we aim to find estimates for $\mathbf{x} = [x(t_0), \ldots, x(t_{N-1})]$ given $\boldsymbol{y}$ and $\boldsymbol{F}$. In GP regression, we assume that $x(t)$ is drawn from a Gaussian process with a kernel $k(t_i, t_j)$ that is parameterized by hyper-parameters $\boldsymbol{\phi}$. For fixed $\boldsymbol{\phi}$, this leads to tractable Gaussian priors over the function $\boldsymbol{x}$ and its derivatives $\dot{\boldsymbol{x}}$. These priors can be combined with the observation models to obtain the generative model summarized in Figure 1.

$$\begin{aligned}
p(\boldsymbol{x} \mid \boldsymbol{\phi}) &= \mathcal{N}(\boldsymbol{x} \mid \mathbf{0}, \boldsymbol{C}_{\boldsymbol{\phi}}) \\
p(\dot{\boldsymbol{x}} \mid \boldsymbol{x}, \boldsymbol{\phi}) &= \mathcal{N}(\dot{\boldsymbol{x}} \mid \boldsymbol{D}\boldsymbol{x}, \boldsymbol{A}) \\
p(\boldsymbol{y} \mid \boldsymbol{x}, \sigma) &= \mathcal{N}(\boldsymbol{y} | \boldsymbol{x}, \sigma^2 \boldsymbol{I}) \\
p(\boldsymbol{F} \mid \dot{\boldsymbol{x}}, \boldsymbol{\gamma}) &= \mathcal{N}(\boldsymbol{F} \mid \dot{\boldsymbol{x}}, \gamma \boldsymbol{I})
\end{aligned}$$

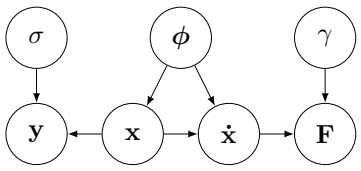


Figure 1: Generative model for Gaussian process regression with derivative observations.

Here, the matrices $\boldsymbol{C}_{\boldsymbol{\phi}}$, $\boldsymbol{D}$ and $\boldsymbol{A}$ are fully determined by the kernel k and the choice of $\boldsymbol{t}$. Since all relevant probability densities are Gaussian, the posteriors can be calculated analytically, as they

will be Gaussian as well. In the rest of this paper, we will focus on the predictive posterior mean and variance for both state and variance, i.e.

$$p(x(\tau)|\boldsymbol{y}, \boldsymbol{F}, \boldsymbol{\phi}, \gamma) = \mathcal{N}(x(\tau)|\mu(\tau), \Sigma(\tau)) \tag{4}$$

and

$$p(x'(\tau)|\boldsymbol{y}, \boldsymbol{F}, \boldsymbol{\phi}, \gamma) = \mathcal{N}(x'(\tau)|\mu'(\tau), \Sigma'(\tau)). \tag{5}$$

$p(\boldsymbol{x}(\boldsymbol{\tau})|\boldsymbol{y}, \boldsymbol{F}, \boldsymbol{\phi}, \gamma)$ These quantities are of particular importance for any task involving smoothing or interpolation. As we shall demonstrate, our scheme allows for exponentially fast decaying error bounds for these quantities as well. For more details regarding notation, please refer to Section C.1.

2.3 Parameter Inference for ODEs with GPs

One important application of this regression technique is parameter inference of differential equations. To simplify notation, we briefly recap the theory for one dimensional systems of ODEs, noting that there is no significant difference to the multidimensional case.

Assume that we are given the noisy observations $\boldsymbol{y}$ of a dynamical system with known parametric form $\dot{x} = f(x, \boldsymbol{\theta})$. The goal is to infer the unknown parameters $\boldsymbol{\theta}$. In ODIN, this equation is now used as a constraint for the probabilistic model in Figure 1. Parameter and states are then found by minimizing the following objective, where $\boldsymbol{F}$ has already been substituted by the constraints:

$$\begin{aligned} \mathcal{R}(\boldsymbol{x},\boldsymbol{F}, \boldsymbol{y}) &= \boldsymbol{x}^T \boldsymbol{C}_{\boldsymbol{\phi}}^{-1} \boldsymbol{x} && (6)\\ &+ (\boldsymbol{x} - \boldsymbol{y})^T \sigma^{-2} (\boldsymbol{x} - \boldsymbol{y}) && (7)\\ &+ (\boldsymbol{F} - \boldsymbol{D}\boldsymbol{x})^T (\boldsymbol{A} + \gamma \boldsymbol{I})^{-1} (\boldsymbol{F} - \boldsymbol{D}\boldsymbol{x}). && (8) \end{aligned}$$

To infer the hyper-parameters from data, Wenk et al. (2020) propose a preprocessing step, in which for each dimension, ODIN maximizes the marginal likelihood by solving

$$\begin{aligned} &\max && p(\boldsymbol{y}|\boldsymbol{\phi}, \sigma) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{\phi}} + \sigma^2 \boldsymbol{I}) && (9)\\ &\text{w.r.t.} && \boldsymbol{\phi}, \sigma && (10) \end{aligned}$$

The model mismatch parameter γ can either be hand-tuned (as done e.g. by Wenk et al. (2018) or Gorbach et al. (2017)) or inferred at run time using ODIN as well. In this case, the risk term of Equations (6) - (8) is extended by the additional summand $\log\det(\boldsymbol{A} + \gamma \boldsymbol{I})$, representing the contribution of the determinant term of the conditional $p(\boldsymbol{F}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\phi}, \gamma)$. In this work, we will refer to this case whenever we speak about learning γ. Clearly, ODIN scales similarly to standard GP regression with derivatives cubically in the amount of observations due to the inversions of $\boldsymbol{C}_{\boldsymbol{\phi}}$ and $\boldsymbol{A}$. Again, all quantities that are not defined in detail can be looked up in Section C.1.

3. QFF for Derivative Information

For standard GP regression, the main computational challenge lies in calculating any terms involving the covariance matrix $\boldsymbol{C}_{\boldsymbol{\phi}}$. In the context of Hermitian quadrature, the work of Mutny and Krause (2018) effectively reduces this complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(NM^2 + M^3)$, where $M << N$ is a fixed constant representing the amount of features used in the feature expansion.

However, as soon as we introduce derivative information, we obtain an additional term depending on the matrices $\boldsymbol{D}$ and $\boldsymbol{A}$. Similar to $\boldsymbol{C}_\phi$, the size of these matrices grows linearly with the amount of observations N, again leading to a computational complexity of $\mathcal{O}(N^3)$. Since the calculation of these matrices involves the first two derivatives of the kernel, we need to extend the standard QFF framework to include such derivatives.

Since the functional form of the kernel dictates the QFF features, any analysis would need to be repeated for every kernel. We will thus restrict ourselves to the RBF kernel, even though it could in principle be extended to any stationary kernel. The RBF kernel is defined as $k(t_i, t_j) = k(t_i - t_j) = k(r_{ij}) = \rho \exp(-\frac{r_{ij}^2}{2l^2})$, where $r_{ij} \coloneqq t_i - t_j$. Here, the variance ρ and lengthscale l are the hyper-parameters of the kernel, which need to be learned in a preprocessing step. The RBF kernel is known for its excellent smoothing properties. Since scaling up the amount of observations is mainly relevant for densely observed trajectories, this makes it an ideal choice for the experiments we will show in Section 5.

3.1 Approximate Derivative of the Kernel

Since both the integral and the derivative operator are linear, we can use their commutativity to obtain new quadrature schemes for the kernel derivatives. For ease of readability, let us define $I(f(\omega)) \coloneqq \int_{-\infty}^{+\infty} e^{-\omega^2} f(\omega) d\omega$ and fix $\rho = \sqrt{\pi}$. This assumption is simply to avoid cluttered notation and will could be lifted by introducing a constant factor instead. For ease of notation, let $r \coloneqq t_i - t_j$.

Applying Bochner's theorem (1) to the RBF kernel with $\rho = \sqrt{\pi}$, we get, similarly to Mutny and Krause (2018),

$$k(t_i, t_j) = I(\cos(\omega r \frac{\sqrt{2}}{l})). \tag{11}$$

Using the linearity of the integral operator, we can differentiate the above equality to obtain

$$\frac{\partial}{t_i} k(t_i, t_j) = \frac{d}{dr} I(\cos(\omega r \frac{\sqrt{2}}{l})) = I(-\frac{\sqrt{2}}{l} \omega \sin(\omega r \frac{\sqrt{2}}{l})). \tag{12}$$

Similarly,

$$\frac{\partial}{\partial t_j} k(t_i, t_j) = I(\omega \frac{\sqrt{2}}{l} \sin(\omega r \frac{\sqrt{2}}{l})) \tag{13}$$

and

$$\frac{\partial^2}{\partial t_i \partial t_j} k(t_i, t_j) = I(\omega^2 \frac{2}{l^2} \cos(\omega r \frac{\sqrt{2}}{l})). \tag{14}$$

These calculations reduce the problem of approximating the kernel derivatives to approximating integrals. Thus, similar to the derivative free case, we can now leverage the powerful framework of Gauss-Hermite quadrature (Hildebrand, 1987).

Let

$$Q_m(f(\omega)) = \sum_{i=1}^{m} W_i^m f(\omega_i^m) \tag{15}$$

denote the Gauss-Hermite quadrature scheme of order m for the function f, where W_i^m are its weights and ω_i^m its abscissas. Construct the $2m$ dimensional feature vector $\boldsymbol{\phi}(x)$ by adding for each i the two components

$$\left[\sqrt{W_i^m}\cos(\omega_i^m\frac{\sqrt{2}}{l}t_i) \quad \sqrt{W_i^m}\sin(\omega_i^m\frac{\sqrt{2}}{l}t_i)\right]. \tag{16}$$

Given this feature vector, we first observe that

$$\boldsymbol{\phi}(t_i)^T\boldsymbol{\phi}(t_j) = \sum_{i=1}^{m} W_i^m\cos(\omega_i^m\frac{\sqrt{2}}{l}r) = Q_m(\cos(\omega\frac{\sqrt{2}}{l}r)), \tag{17}$$

as desired when recovering the QFF approximation for $k(t_i, t_j)$ for calculating $\boldsymbol{C_\phi}$.

However, this feature expansion can also be differentiated w.r.t. t_i to obtain

$$\boldsymbol{\phi}(t_i)' = \begin{bmatrix} -\sqrt{W_i^m}\frac{\sqrt{2}}{l}\omega_i^m\sin(\omega_i^m\frac{\sqrt{2}}{l}t_i) \\ \sqrt{W_i^m}\frac{\sqrt{2}}{l}\omega_i^m\cos(\omega_i^m\frac{\sqrt{2}}{l}t_i)\end{bmatrix}_{i=1\ldots m}. \tag{18}$$

Using trigonometric identities, it can be shown that this feature expansion indeed yields

$$\boldsymbol{\phi}(t_i)'^T\boldsymbol{\phi}(t_j) = Q_m(-\frac{\sqrt{2}}{l}\omega\sin(\omega r\frac{\sqrt{2}}{l})), \tag{19}$$

$$\boldsymbol{\phi}(t_i)^T\boldsymbol{\phi}(t_j)' = Q_m(\frac{\sqrt{2}}{l}\omega\sin(\omega r\frac{\sqrt{2}}{l})), \tag{20}$$

$$\boldsymbol{\phi}(t_i)'^T\boldsymbol{\phi}(t_j)' = Q_m(\omega^2\frac{2}{l^2}\cos(\omega r\frac{\sqrt{2}}{l})). \tag{21}$$

Thus, the features given by (18) represent the quadrature schemes for the integrals shown in Equations (12) and (14). Thus, the kernel derivatives involved in calculating $\boldsymbol{A}$ and $\boldsymbol{D}$ can be approximated by

$$\frac{\partial}{\partial x}k(x,y) \approx \boldsymbol{\phi}(x)'^T\boldsymbol{\phi}(y) \tag{22}$$

and

$$\frac{\partial^2}{\partial x\partial y}k(x,y) \approx \boldsymbol{\phi}(x)'^T\boldsymbol{\phi}(y)'. \tag{23}$$

As we shall prove and empirically validate in the rest of this paper, this feature expansion is efficient in the sense that the approximation error decays *exponentially* in the amount of features. This allows us to obtain accurate approximations of the kernel derivatives even for a small amount of *deterministically* chosen features.

3.2 Application to GPR with Derivatives

To apply this approximation to GPR with derivatives, we make use of the matrix inversion lemma.

Lemma 1 (Matrix Inversion Lemma) *Let A and C be invertible matrices. Then for any matrices U, V of appropriate dimensions, it holds that* $(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}$.

In the context of standard GP regression (e.g. Rahimi and Recht, 2008; Mutny and Krause, 2018), this can be used as follows. Let $\boldsymbol{\Phi} \in \mathbb{R}^{2m \times N}$ be such that its columns are given by the feature vectors $\boldsymbol{\phi}(t_i)$ as defined in (16) for a quadrature scheme of order m. After adding a small jitter λ on the diagonal of $\boldsymbol{C}_{\boldsymbol{\phi}}$, we can use Lemma 1 to obtain

$$(\boldsymbol{C}_{\boldsymbol{\phi}} + \lambda \boldsymbol{I})^{-1} \approx \frac{1}{\lambda}(\boldsymbol{I} - \boldsymbol{\Phi}^T(\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \lambda \mathbb{I})^{-1}\boldsymbol{\Phi}) \tag{24}$$

Here, $\approx$ should be read as an approximation up to an exponentially decaying error in M, which we will quantify in Section 4.

Since all matrices required to calculate the quantities in Equations (4) and (5) allow for a direct feature expansion, the matrix inversion lemma can be directly applied to approximate these quantities. Since we work directly with the posterior mean and variance, we do not even need a jitter term due to the presence of the noise terms.

3.3 SLEIPNIR

However, extending this concept to Equation (8) is slightly more tricky, as neither $\boldsymbol{D}$ nor $\boldsymbol{A}$ allow for a direct feature expansion. Nevertheless, it is still possible to derive a scalable approximation using the previously introduced feature expansions:

First, start by collecting the appropriate feature vectors from Equations (16) and (18) to write $\boldsymbol{C}'_{\boldsymbol{\phi}} \approx \boldsymbol{\Phi}'^T\boldsymbol{\Phi}$, $'\boldsymbol{C}_{\boldsymbol{\phi}} \approx \boldsymbol{\Phi}^T\boldsymbol{\Phi}'$ and $\boldsymbol{C}''_{\boldsymbol{\phi}} \approx \boldsymbol{\Phi}'^T\boldsymbol{\Phi}'$. These approximations can then be inserted into the term given by Equation (8). By applying Theorem 1 multiple times, it is possible to finally obtain

$$\begin{aligned} \boldsymbol{z}^T(\boldsymbol{A} + \gamma \boldsymbol{I})^{-1}\boldsymbol{z} \approx \\ \frac{1}{\gamma}\boldsymbol{z}^T(\mathbb{I} - \boldsymbol{\Phi}'^T(\boldsymbol{\Phi}'\boldsymbol{\Phi}'^T + \frac{\gamma}{\lambda}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \gamma\mathbb{I})^{-1}\boldsymbol{\Phi}')\boldsymbol{z}, \end{aligned} \tag{25}$$

where

$$\boldsymbol{z} \coloneqq \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{x}) - \boldsymbol{D}\boldsymbol{x} \approx \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{x}) - \boldsymbol{\Phi}'^T(\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \lambda\mathbb{I})^{-1}\boldsymbol{\Phi}\boldsymbol{x}.$$

Combining these approximations with the one given by Equation (24), we obtain a computationally efficient scheme for calculating all objective functions in ODIN. Furthermore, both the additional term $\log\det(\boldsymbol{A} + \gamma\boldsymbol{I})$ obtained when learning γ and learning the hyperparameters via Equation (9) can be scaled similarly using these approximations and directly applying the matrix inversion lemma.

Overall, this means that the original complexity of $\mathcal{O}(N^3)$ has been successfully reduced to $\mathcal{O}(NM^2 + M^3)$. For $M < N$, this is significantly accelerating ODIN, which is why we name our approximation scheme SLEIPNIR. The resulting algorithm will be referred to as ODIN-S.

4. Theoretical Results

It is clear that this bound is especially appealing, since the error will exhibit exponential decay for $m > l^2$. In this section, we state that this behavior can be generalized to our approximations of the kernel derivatives. As we have demonstrated in the previous section, our feature approximation can be used to efficiently reduce the computational complexity of GPRD and ODIN. In this section, we will provide an in-depth error analysis of this approximation, demonstrating the key theoretical result of this paper: The favorable properties of standard QFF carry over to the derivative case and can be efficiently deployed in practically relevant algorithms.

4.1 Exponentially Decaying Kernel Approximation Errors

Let us define

$$E_m := \sqrt{\pi}\frac{1}{m^m}\left(\frac{e}{4l^2}\right)^m \tag{26}$$

In the context of standard GP regression, Mutny and Krause (2018) have shown that

$$|k(t_i, t_j) - \boldsymbol{\phi}(t_i)^T \boldsymbol{\phi}(t_j)| \leq E_m \tag{27}$$

The results are summarized in Theorem 2, with a detailed proof given in the appendix in Section A.

Theorem 2 *Let $k(t_i, t_j)$ be defined as in Equation* (11) *and consider the Gauss-Hermite quadrature scheme of order m as described by Equation* (15)*, defining $\boldsymbol{\phi}(t_i)$ and $\boldsymbol{\phi}(t_i)'$ as in Equations* (16) *and* (18)*. Then, for $|r| = |t_i - t_j| \leq 1$, it holds that*

(i) $|\frac{\partial}{\partial t_i} k(t_i, t_j) - \boldsymbol{\phi}(t_i)'^T \boldsymbol{\phi}(t_j)| \leq \frac{2e}{l^2} E_{m-2}$

(ii) $|\frac{\partial^2}{\partial t_i \partial t_j} k(t_i, t_j) - \boldsymbol{\phi}(t_i)'^T \boldsymbol{\phi}(t_j)'| \leq \frac{2e}{l^4} E_{m-3}$

where E_m is defined as in Equation (26)

Not that in the above Theorem, $|r| < 1$ can always be achieved by rescaling the data and adapting the lengthscale l accordingly.

While the error is slightly larger for the derivative approximations than for the kernel itself, it is important to note that we are still getting the same exponential decaying behavior.

4.2 GPR with Derivatives Bounds

As we shall see in this section, this exponential decay carries over nicely to the case of GPR with derivatives. Define $e_{\tilde{\mu}}$, $e_{\tilde{\Sigma}}$, $e_{\tilde{\mu}'}$ and $e_{\tilde{\Sigma}'}$ as the absolute error between the feature approximations and the corresponding accurate quantities of the means and covariances of Equations (4) and (5). For each $\tau \in \mathbb{R}$, define e_{tot} as the maximum of these four errors. Using these definitions, we can show the exponential relation between feature approximation order m and the corresponding approximation error, as summarized in Theorem 3, with proof in the appendix.

Theorem 3 *Let us consider an RBF kernel with hyperparameters (ρ, l) and domain $[0, 1]$. Define $c := \min(\gamma, \sigma^2)$ and $R := \max(||\boldsymbol{y}||_\infty, ||\boldsymbol{F}||_\infty)$. Let $C > 0$. Let us consider a QFF approximation scheme of order $m \geq 3 + \max\left(\frac{e}{2l^2}, \log\left(\frac{270n^2\rho^3 R}{l^8 c^2 C}\right)\right)$. Then, it holds for all $\tau \in [0, 1]$ that $e_{tot} \leq C$.*

From this theorem, we can also observe the following fact: If we decrease the acceptable worst case performance C, we clearly need more features. However, due to the logarithm in the theorem, an *exponential* decay in C only leads to *linear* growth in m, all other things being equal.

4.3 SLEIPNIR Bounds

Similarly, we can observe that when applying SLEIPNIR to ODIN for a fixed feature vector length M, we will always create a small error in the objective function. However, the deterministic nature of the QFF approximation still allows us to choose the amount of features in a way such that the approximation error will always be guaranteed to be smaller than a pre-chosen threshold. This result is summarized in Theorem 4 and proven in the appendix, Section E.

Theorem 4 *Let $\mathcal{R}$ be the* ODIN*-objective as defined in Equations* (6)-(8) *and let $\tilde{\mathcal{R}}$ denote its counterpart obtained by approximating the matrices $\boldsymbol{C}_{\boldsymbol{\phi}}$, $\boldsymbol{A}$ and $\boldsymbol{D}$ as described in Section 3.3. Assume the parameters λ, γ, $\boldsymbol{\phi} = (\rho, l)$ and N to be fixed. Suppose $N \geq 60$ and let $1 > \epsilon > 0$. Let m denote the order of the quadrature scheme. Then*

$$m \geq 10 + \max\{\frac{e}{2l^2}, \log_2(\frac{\rho^2 n^3}{\lambda^2 \gamma l^4 \epsilon})\} \tag{28}$$

implies

$$\frac{|R_{\lambda\gamma\phi}(x,\theta) - \tilde{R}_{\lambda\gamma\phi}(x,\theta)|}{R_{\lambda\gamma\phi}(x,\theta)} \leq \epsilon \tag{29}$$

for any configuration of the variables $\boldsymbol{x}$ and $\boldsymbol{\theta}$.

Again, it should be noted that the threshold ϵ appears inside the logarithm. Thus, an *exponential* decrease of the allowed error only requires a *linear* increase in the amount of features.

In this theorem, the bound on N is purely aesthetically motivated, to reduce the amount of terms. From a practical perspective however, one would probably want to use ODIN without any approximations anyways if only < 60 observations are available. Furthermore, the bound on ϵ just requires that investigating an approximation scheme with more than 100% relative error is of little interest.

5. Empirical Validation

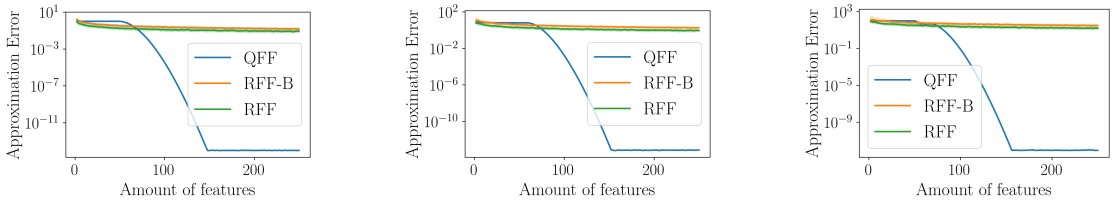
In this section, we will provide empirical validation for all theoretical claims made in this paper. We will start by showing the exponentially decaying behavior of the kernel approximation error in Section 5.1. In Sections 5.2 and 5.3, we show that this exponentially decaying behavior carries over to both GPR with derivatives and ODIN. Ultimately, we will conclude by showing how ODIN with SLEIPNIR can be used on a realistic dataset, demonstrating its pratical applicability in Section 5.4.

Wherever applicable, we will use three standard benchmark systems from the GP-based gradient matching literature, namely the Lotka Volterra (LV) system (Lotka, 1932), the Protein Transduction (PT) system (Vyshemirsky and Girolami, 2007) and the Lorenz system (Lorenz, 1963). While the LV system is quite easy to fit, both PT and Lorenz offer interesting challenges. The non-stationary dynamics of PT make it a formidable challenge for collocation methods. Classic literature (e.g. Dondelinger et al., 2013; Gorbach et al., 2017; Wenk et al., 2020) sidestep this issue by using a non-stationary sigmoid kernel. However, as we shall demonstrate in this section, this is not a problem in the case of densely observed trajectories, since we will be using the RBF kernel in all experiments. Finally, the Lorenz system is interesting due to its chaotic behavior. Chaotic systems are an interesting challenge for many parameter inference schemes due to the potentially high sensitivity to parameter changes and the presence of many local optima. For a full description of the experimental setup

and all metrics used please refer to Section F. For all experiments, we created our data-set using numerical integration of the ground truth parameters. We then added 25 different noise realizations to obtain 25 different data sets. This allows us to quantify robustness w.r.t. noise by showing median as well as 20% and 80% quantiles over these noise realizations for each experiment. In all experiments, we trained γ and learned the kernel hyperparameters from the data using the scalable approximations described in the previous section. To compare with ODIN-S, we chose to combine ODIN with the RFF and RFF-B feature expansions detailed in Section 2.

5.1 Exponentially Decaying Kernel Approximation Errors

In Figure 2, we evaluate the maximum error of the feature approximations over an interval $r \in [0, 1]$. From Theorem 2, we would expect an exponential decay in the approximation error for both the kernel and the first and second order derivatives. This exponential decay is clearly visible and continues until we hit numerical boundaries. In Figure 2, the l was chosen to be 0.1. However, as can be seen on the additional plots in the appendix (Sec. B), this behavior is robust across different lengthscales. Figure 2 allows for the interesting observation that next to the exponential decay of QFF, the error of the RFF almost looks constant, even though it decays linearly.



(a) $k(r)$ (b) $k'(r)$ (c) $k''(r)$

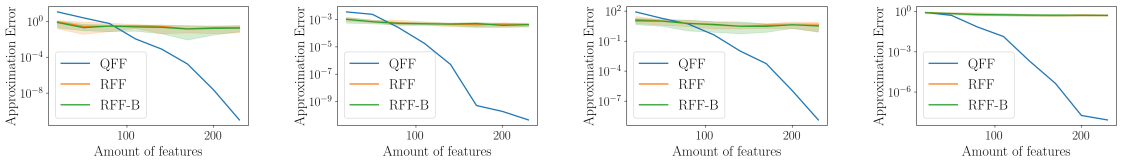
Figure 2: Comparing the maximum error of different feature expansions over $r \in [0, 1]$. For the random feature expansions, we show median as well as 12.5% and 87.5% quantiles over 100 random samples. Due to the exponential decay of the error of the QFF approximation, this stochasticity is barely visible. As given by the theoretical analysis, the error is a bit higher for the derivatives, but still decaying exponentially.

5.2 GPR with Derivatives

A similar behavior can be observed for the absolute error of the GP posterior we introduced in Section 2.2. As can be seen in Figure 3, the exponential decay of the error predicted in Section 4.2 also appears in practice. While we only show one state of one experiment, this behavior is consistent across experiments and noise settings, as can be seen by looking at the additional plots presented in the appendix, Section D.

5.3 ODIN-S on Standard Benchmark Systems

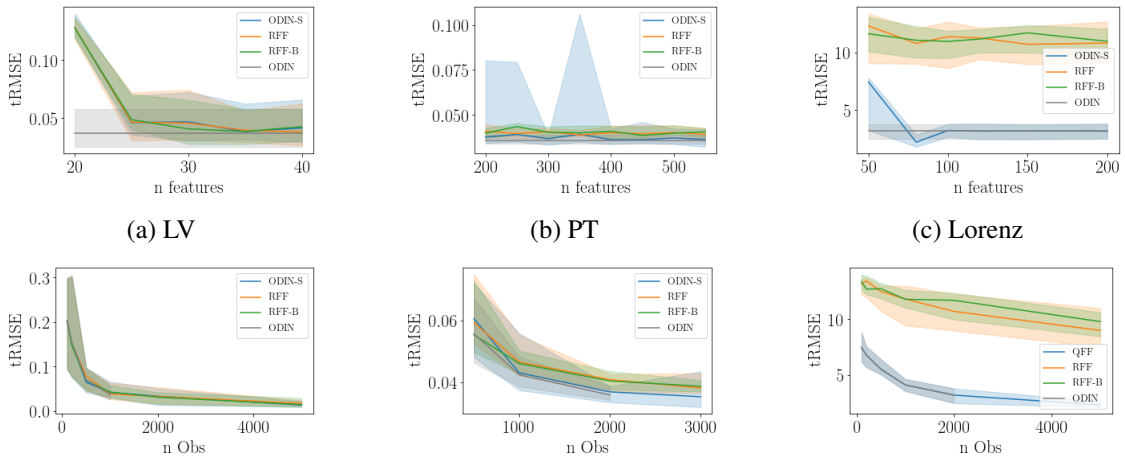
In Figure 4, we compare the performance of ODIN-S against accurate ODIN as well as ODIN augmented with RFF and RFF-B on three standard benchmark systems. In the top row, we keep the total amount of observations fixed to 1000 for LV and 2000 for PT and Lorenz, while varying the length of the feature vector. In the bottom row, we keep the amount of features fixed to 40 for LV, 300 for PT and 150 for Lorenz. All the data was created using observation noise with $\sigma^2 = 0.1$ for



(a) μ_0 (b) Σ_0 (c) μ_0' (d) Σ_0'

Figure 3: Approximation error of the different feature approximations compared to the accurate GP, evaluated at $t = 0.8$ for the Lorenz system with 1000 observations and an SNR of 100. For each feature, we show the median as well as the 12.5% and 87.5% quantiles over 10 independent noise realizations for the first state dimension.

LV, $\sigma^2 = 0.01$ for PT and a signal-to-noise ratio of 5 for Lorenz. Due to computational restrictions, it was not possible to evaluate accurate ODIN beyond what is shown in the plots. To provide an idea of the robustness of the evaluation, different noise, feature and observation settings are investigated in the appendix, Section G.



(a) LV (b) PT (c) Lorenz

Figure 4: Comparing the tRMSE of the ODE parameters obtained via the three different approximation schemes. In the top row, we show how the results obtained by ODIN-S converge to the accurate results when increasing the amount of features. In the second row, we show the learning curves for a fixed feature vector length, demonstrating that increasing the amount of observations actually improves the parameter estimates and decreases the tRMSE.

As expected by our theoretical analysis, the trajectory RMSE of the SLEIPNIR-based ODIN-S eventually converges to the approximation-free ODIN in both median and quantiles if we increase the number of Fourier features. This clearly validates the main result of Theorem 4, expecting an exponentially small error. Also, it is interesting to observe that the MCMC-based RFF and RFF-B seem to struggle to keep up with the approximation quality of SLEIPNIR. While there is not much difference visible for the smooth and simple Lotka Volterra system, this is clearly visible for PT and especially striking for the chaotic Lorenz system. Complementing the results of Figure 2 and Figure 3, this further illustrates the power of the SLEIPNIR approximation, especially in this setting, where the kernel inputs are one-dimensional.

Table 1: Comparing the run time of accurate ODIN against the run time of the feature approximations per iteration in miliseconds. For each setting, we show the median $\pm$ on standard deviation over 15 different iterations. As expected, the feature approximations lead to order of magnitude reductions in run time.

	accurate	feature
LV	973 ± 42.5	1.24 ± 0.243
PT	17900 ± 668	133 ± 12.8
Lorenz	10700 ± 466	13.2 ± 0.566

In Table 1, we finally compare the run time per iteration of ODIN with ODIN-S. Since the run time only depends on the amount of features and not on how the features were obtained, we omitted RFF and RFF-B. The run time was evaluated using the same amount of observation and amount of feature combinations as in Figure 4, namely $(2000, 150)$ for LV with $\sigma^2 = 0.1$, $(2000, 300)$ for PT with $\sigma^2 = 0.01$ and $(1000, 40)$ for Lorenz with an SNR of 5. While it should already be clear from theoretical analysis that ODIN-S scales linearly in the amount of observations N and cubic in the length of the feature vector M, this was confirmed empirically as well. However, due to space restrictions, the plots have been moved to the appendix and are shown in Figure 29 and Figure 30.

5.4 Practical Applicability

In a final experiment, we show that ODIN-S is able to scale to realistic data sets. For this, we introduce a 12-dimensional ODE system representing the dynamics of a 6DOF quadrocopter. We observe the system under Gaussian noise with SNR=10 over the time interval $t = [0, 15]$. We assume a sampling frequency of 1kHz, leading to 15'000 observations. We then run ODIN-S on a standard laptop (Lenovo Carbon X1) and obtain results in roughly 80min. Up to our knowledge, this is the first time that a system of such dimensions has been solved with a Gaussian process based parameter inference scheme, clearly demonstrating the power of our framework. The resulting trajectories are shown in Figure 5, including example observation points to visualize the noise level. The estimates of ODIN-S are so good that the ground truth is barely visible.

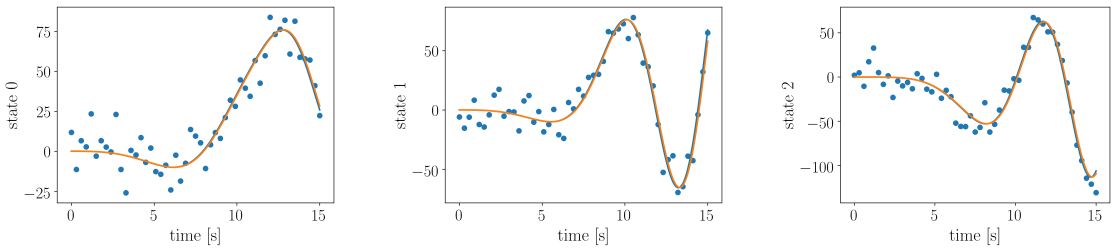


Figure 5: State trajectories of the first three states obtained by integrating the parameters inferred by ODIN-S (orange). The blue line represents the ground truth, while the blue dots show every 300-th observation for a signal-to-noise ratio of 10. States 5-12 have been moved to the appendix, Section G.3.

6. Conclusion

In this work, we introduced a new theoretical framework to scale up Gaussian process regression with derivative information, extending existing work based on quadrature Fourier features. We derived and proved deterministic, exponentially fast decaying error bounds for approximating the derivatives of an RBF kernel. We then combined these insights to create a computationally efficient approximation scheme for both standard GP regression with derivatives as well as the parameter inference scheme ODIN. The theoretical analysis of this approximation yielded deterministic, non-asymptotic error bounds. In an extensive empirical evaluation, we then showed orders of magnitude improvements on the run time without sacrificing accuracy. In future work, we are excited to see how SLEIPNIR could be deployed in other areas such as scaling up Bayesian optimization with derivatives (e.g. Wu et al., 2017) or probabilistic numerics (Hennig et al., 2015).

Acknowledgments

This research was supported by the Max Planck ETH Center for Learning Systems. This projecthas received funding from the European Research Council(ERC) under the European Union's Horizon 2020 researchand innovation programme grant agreement No 815943.

References

Gabriele Abbati, Philippe Wenk, Michael A. Osborne, Andreas Krause, Bernhard Schölkopf, and Stefan Bauer. AReS and MaRS adversarial and MMD-minimizing regression for SDEs. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1–10, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Ben Calderhead, Mark Girolami, and Neil D Lawrence. Accelerating bayesian inference over nonlinear differential equations with gaussian processes. In *Advances in neural information processing systems*, pages 217–224, 2009.

John P Cunningham, Krishna V Shenoy, and Maneesh Sahani. Fast gaussian process methods for point process intensity estimation. In *Proceedings of the 25th international conference on Machine learning*, pages 192–199. ACM, 2008.

Frank Dondelinger, Dirk Husmeier, Simon Rogers, and Maurizio Filippone. Ode parameter inference using adaptive gradient matching with gaussian processes. In *Artificial intelligence and statistics*, pages 216–228, 2013.

Leander Dony, Fei He, and Michael PH Stumpf. Parametric and non-parametric gradient matching for network inference: a comparison. *BMC bioinformatics*, 20(1):52, 2019.

David Eriksson, Kun Dong, Eric Lee, David Bindel, and Andrew G Wilson. Scaling gaussian process regression with derivatives. In *Advances in Neural Information Processing Systems*, pages 6867–6877, 2018.

Nico S Gorbach, Stefan Bauer, and Joachim M Buhmann. Scalable variational inference for dynamical systems. In *Advances in Neural Information Processing Systems*, pages 4806–4815, 2017.

Philipp Hennig, Michael A Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150142, 2015.

James Hensman, Nicolas Durrande, Arno Solin, et al. Variational fourier features for gaussian processes. *Journal of Machine Learning Research*, 18(151):1–151, 2017.

Francis Begnaud Hildebrand. *Introduction to numerical analysis*. Courier Corporation, 1987.

Miguel Lazaro-Gredilla, Joaquin Quiñonero-Candela, Carl Edward Rasmussen, and Anibal R Figueiras-Vidal. Sparse spectrum gaussian process regression. *Journal of Machine Learning Research*, 11(Jun):1865–1881, 2010.

Edward N Lorenz. Deterministic nonperiodic flow. *Journal of the atmospheric sciences*, 20(2): 130–141, 1963.

Alfred J Lotka. The growth of mixed populations: two species competing for a common food supply. *Journal of the Washington Academy of Sciences*, 22(16/17):461–469, 1932.

Benn Macdonald and Dirk Husmeier. Gradient matching methods for computational inference in mechanistic models for systems biology: a review and comparative analysis. *Frontiers in bioengineering and biotechnology*, 3:180, 2015.

Jonas Mockus. *Bayesian approach to global optimization: theory and applications*, volume 37. Springer Science & Business Media, 2012.

Mojmir Mutny and Andreas Krause. Efficient high dimensional bayesian optimization with additivity and quadrature fourier features. In *Advances in Neural Information Processing Systems*, pages 9005–9016, 2018.

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

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

CE. Rasmussen and CKI. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, January 2006.

W. Rudin. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1976. ISBN 9780070856134.

Simo Sarkka, Arno Solin, and Jouni Hartikainen. Spatiotemporal learning via infinite-dimensional bayesian filtering and smoothing: A look at gaussian process regression through kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013.

Sambu Seo, Marko Wallat, Thore Graepel, and Klaus Obermayer. Gaussian process regression: Active data selection and test point rejection. In *Mustererkennung 2000*, pages 27–34. Springer, 2000.

Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264, 2006.

Ercan Solak, Roderick Murray-Smith, William E Leithead, Douglas J Leith, and Carl E Rasmussen. Derivative observations in gaussian process models of dynamic systems. In *Advances in neural information processing systems*, pages 1057–1064, 2003.

Arno Solin and Simo Särkkä. Hilbert space methods for reduced-rank gaussian process regression. *Statistics and Computing*, pages 1–28.

Arno Solin, Manon Kok, Niklas Wahlström, Thomas B Schön, and Simo Särkkä. Modeling and interpolation of the ambient magnetic field by gaussian processes. *IEEE Transactions on robotics*, 34(4):1112–1127, 2018.

Zoltán Szabó and Bharath K Sriperumbudur. On kernel derivative approximation with random fourier features. *arXiv preprint arXiv:1810.05207*, 2018.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

Vladislav Vyshemirsky and Mark A Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833–839, 2007.

Philippe Wenk, Alkis Gotovos, Stefan Bauer, Nico Gorbach, Andreas Krause, and Joachim M Buhmann. Fast gaussian process based gradient matching for parameter identification in systems of nonlinear odes. *arXiv preprint arXiv:1804.04378*, 2018.

Philippe Wenk, Gabriele Abbati, Stefan Bauer, Michael A Osborne, Andreas Krause, and Bernhard Schölkopf. Odin: Ode-informed regression for parameter and state inference in time-continuous dynamical systems. 2020.

Christopher KI Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688, 2001.

Andrew Wilson and Hannes Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *International Conference on Machine Learning*, pages 1775–1784, 2015.

Andrew G Wilson, Elad Gilboa, Arye Nehorai, and John P Cunningham. Fast kernel learning for multidimensional pattern extrapolation. In *Advances in Neural Information Processing Systems*, pages 3626–3634, 2014.

Jian Wu, Matthias Poloczek, Andrew G Wilson, and Peter Frazier. Bayesian optimization with gradients. In *Advances in Neural Information Processing Systems*, pages 5267–5278, 2017.

Appendix A. Kernel Approximation Error Bounds

For any function f, let $I(f(\omega)) \coloneqq \int_{-\infty}^{+\infty} e^{-\omega^2} f(\omega) d\omega$.

For any $r \in [0,1]$, let $k(r) \coloneqq \sqrt{\pi} e^{-\frac{r^2}{2l^2}} = \int_{-\infty}^{+\infty} e^{-\omega^2} cos(\omega r \frac{\sqrt{2}}{l}) d\omega = I(cos(\omega r \frac{\sqrt{2}}{l}))$.
Let $Q_m(f) = \sum_{i=1}^{m} W_i^m f(x_i^m)$ denote the Gauss-Hermite quadrature scheme of order m and function f with weights $W_i^m \geq 0$ and abscissas x_i^m and let $S^m \coloneqq \{x_1^m, x_2^m, ..., x_m^m\}$ denote the set of these abscissas.
From Gauss-Hermite quadrature, we know: If f is a polynomial of degree $2m-1$ at most, then $Q_m(f(\omega)) = I(f(\omega))$, i.e the quadrature scheme exactly computes the integral.

Let $H_m(x)$ be the Hermite polynomial of order m and $h_m(x) \coloneqq \frac{H_m(x)}{2^m}$ be its normalized version.
Since $I(H_i(\omega)H_j(\omega)) = r_{ij} 2^j j! \sqrt{\pi}$, we know that $I(h_m(\omega)^2) = \sqrt{\pi} \frac{m!}{2^m}$.

Let $E_m \coloneqq \sqrt{\pi} \frac{1}{m^m} (\frac{e}{4l^2})^m$.
Using this definition, we can restate the error bounds derived by Mutny and Krause (2018) as

$$|I(\cos(\omega r \frac{\sqrt{2}}{l})) - Q_m(\cos(\omega r \frac{\sqrt{2}}{l}))| \leq E_m \tag{30}$$

and our bounds from Theorem 2 as

$$\frac{\sqrt{2}}{l} |I(\omega \sin(\omega r \frac{\sqrt{2}}{l})) - Q_m(\omega \sin(\omega r \frac{\sqrt{2}}{l}))| \leq 8(m-1)E_{m-1} \leq \frac{2e}{l^2} E_{m-2}, \tag{31}$$

$$\frac{2}{l^2} |I(\omega^2 \cos(\omega r \frac{\sqrt{2}}{l})) - Q_m(\omega^2 \cos(\omega r \frac{\sqrt{2}}{l}))| \leq \frac{4}{l^2}(m-1)E_{m-2} \leq \frac{2e}{l^4} E_{m-3}. \tag{32}$$

Similar to normal quadrature, our proof technique is based on choosing polynomial q of degree $2m-1$ whose values and derivatives agree with the original function f and its derivatives at a specific set of points. We can then define the remainder $s \coloneqq f - q$. If we now approximate the inegral of the function f by the integral of the polynomial q, bounding the approximation error is equivalent to bounding the integral of the remainder r.

We thus need two important components. After showing the existence of polynomials of a required degree that agree with f at specific points, we need some results regarding the residuals of polynomial approximations to functions. These components are demonstrated in the next two sections, which will then be combined in the final proof in Section A.3.

A.1 Polynomial approximation residuals

First, let us restate the following two well known Lemmata (see e.g. Hildebrand, 1987):

Lemma 5 *Let f be a real function n times continuously differentiable and $q_n(x)$ a polynomial of degree $n-1$ that agrees with f at the set of distinct points $S = \{x_1, x_2, ..., x_n\}$. Let $\pi(x) = \prod_{i=1}^{n}(x - x_i)$. Then $\forall x \in \mathbb{R}$ we have*

$$f(x) - q_n(x) = \frac{f^{(n)}(\xi)}{n!} \pi(x) \tag{33}$$

for some $\xi = \xi(x) \in I$, where I is the interval limited by the smallest and largest of the numbers $x_1, x_2, ..., x_n$ and x.

Lemma 6 *Let f be a real function $2n$ times continuously differentiable and $q_{2n}(x)$ a polynomial of degree $2n-1$ that agrees with f at the set of distinct points $S = \{x_1, x_2, ..., x_n\}$ and its derivative also agrees with f' at S. Let $\pi(x) = \prod_{i=1}^{n}(x - x_i)^2$. Then $\forall x \in \mathbb{R}$ we have*

$$f(x) - q_{2n}(x) = \frac{f^{(2n)}(\xi)}{(2n)!}\pi(x) \tag{34}$$

for some $\xi = \xi(x) \in I$, where I is the interval limited by the smallest and largest of the numbers $x_1, x_2, ..., x_n$ and x.

For our proof, we will need the following two extensions:

Lemma 7 *Let f be a real function $2n+1$ times continuously differentiable and $q_{2n+1}(x)$ a polynomial of degree $2n$ that agrees with f at the set of distinct points $S = \{x_1, x_2, ..., x_n, x_*\}$ and its derivative also agrees with f' at $S \setminus \{x_*\}$. Let $\pi(x) = (x - x_*)\prod_{i=1}^{n}(x - x_i)^2$. Then $\forall x \in \mathbb{R}$ we have*

$$f(x) - q_{2n+1}(x) = \frac{f^{(2n+1)}(\xi)}{(2n+1)!}\pi(x) \tag{35}$$

for some $\xi = \xi(x) \in I$, where I is the interval limited by the smallest and largest of the numbers $x_1, x_2, ..., x_n, x_*$ and x.

Proof 1 *If $x \in S$ the property holds. Otherwise, fix $\bar{x}$. We define $F(x) = f(x) - q_{2n+1}(x) - K\pi(x)$, where K is chosen such that $F(\bar{x}) = 0$. We see that F has $n+2$ distinct roots at I, so by Rolle's theorem F' has $n+1$ distinct roots in I which are different from the points of S (because F' vanishes at intermediate points). Moreover, F' vanishes at all the points in $S \setminus \{x_*\}$ so in total F' has at least $2n+1$ distinct roots in I. Consequently, by Rolle's theorem again, F'' has $2n$ distinct roots in I, ..., $F^{(2n+1)}$ has one root $\xi \in I$. At this point we have*

$$0 = F^{(2n+1)}(\xi) = f^{(2n+1)}(\xi) - (2n+1)!K \Leftrightarrow K = \frac{f^{(2n+1)}(\xi)}{(2n+1)!}. \tag{36}$$

So $F(\bar{x}) = 0 = f(\bar{x}) - q_{2n+1}(\bar{x}) - \frac{f^{(2n+1)}(\xi)}{(2n+1)!}\pi(\bar{x})^2 \Leftrightarrow f(\bar{x}) - q_{2n+1}(\bar{x}) = \frac{f^{(2n+1)}(\xi)}{(2n+1)!}\pi(\bar{x})^2$. And since $\bar{x}$ was chosen arbritrarily the result holds for all x.

Lemma 8 *Let f be a real function $2n+1$ times continuously differentiable and $q_{2n+1}(x)$ a polynomial of degree $2n$ that agrees with f at the set of distinct points $S = \{x_1, x_2, ..., x_n\}$, its derivative also agrees with f' at S and its second derivative agrees with the second derivative of f at x_1. Let $\pi(x) = (x - x_1)\prod_{i=1}^{n}(x - x_i)^2$. Then $\forall x \in \mathbb{R}$ we have*

$$f(x) - q_{2n+1}(x) = \frac{f^{(2n+1)}(\xi)}{(2n+1)!}\pi(x) \tag{37}$$

for some $\xi = \xi(x) \in I$, where I is the interval limited by the smallest and largest of the numbers $x_1, x_2, ..., x_n$ and x.

Proof 2 *If $x \in S$ the property holds. Otherwise, fix $\bar{x}$. We define $F(x) = f(x) - q_{2n+1}(x) - K\pi(x)$, where K is chosen such that $F(\bar{x}) = 0$. We see that F has $n+1$ distinct roots at I, so by Rolle's*

theorem F' has n distinct roots in I which are different from the points of S (because F' vanishes at intermediate points). Moreover, F' vanishes at all the points in S so in total F' has at least $2n$ distinct roots in I. Consequently, by Rolle's theorem again, F'' has $2n-1$ distinct roots in I which are different from the points of S (because F'' vanishes at intermediate points) and also F'' vanishes at x_1, so in total F'' vanishes at $2n$ points (at least), ..., $F^{(2n+1)}$ has one root $\xi \in I$. At this point we have

$$0 = F^{(2n+1)}(\xi) = f^{(2n+1)}(\xi) - (2n+1)!K \Leftrightarrow K = \frac{f^{(2n+1)}(\xi)}{(2n+1)!}. \tag{38}$$

So $F(\bar{x}) = 0 = f(\bar{x}) - q_{2n+1}(\bar{x}) - \frac{f^{(2n+1)}(\xi)}{(2n+1)!}\pi(\bar{x})^2 \Leftrightarrow f(\bar{x}) - q_{2n+1}(\bar{x}) = \frac{f^{(2n+1)}(\xi)}{(2n+1)!}\pi(\bar{x})^2$. And since $\bar{x}$ was chosen arbritrarily the result holds for all x.

A.2 Existence of polynomials that agree with function at certain points

Lemma 5 - 8 assume the existence of polynomials with certain properties. In the following, we will show that such polynomials actually exist and how they can be constructed.

Let $S = \{x_1, x_2, \cdots x_n\}$ be a set of n distinct points in $\mathbb{R}$ and define $\pi(x) = \prod_{i=1}^{n}(x - x_i)$ and for $i = 1, ..., n$, $l_i(x) = \frac{\pi(x)}{(x-x_i)\pi'(x_i)}$. It holds that $l_i(x_j) = r_{ij}$ and that $deg(l_i) = n - 1$. Moreover, we define for $i = 1, ..., n$, $h_i(x) = l_i^2(x)(-2l_i'(x_i)(x - x_i) + 1)$ and $\bar{h}_i(x) = l_i^2(x)(x - x_i)$. It holds that $h_i(x_j) = r_{ij}$, $h_i'(x_j) = 0$, $\bar{h}_i(x_j) = 0$ and $\bar{h}_i'(x_j) = r_{ij}$ and $deg(h_i) = deg(\bar{h}_i) = 2n - 1$ for $i = 1, ..., n$. Thus, we can directly state the following four Lemmata

Lemma 9 *Let f be a real function and define $y_1(x) = \sum_{i=1}^{n} f(x_i)l_i(x)$. Then y_1 is a polynomial with degree $n - 1$ that agrees with f at S.*

Lemma 10 *Let f be a real differential function and define $y_2(x) = \sum_{i=1}^{n} f(x_i)h_i(x) + \sum_{i=1}^{n} f'(x_i)\bar{h}_i(x)$. Then y_2 is a polynomial with degree $2n - 1$ that agrees with f at S and its derivative also agrees with the derivative in our proofs of f at S.*

Lemma 11 *Consider a point $x_0 \notin S$. Let f be a real differential function and define $y_3(x) = y_2(x) + \frac{f(x_0)-y_2(x_0)}{\pi^2(x_0)}\pi^2(x)$. Then y_3 is a polynomial with degree $2n$ that agrees with f at $S \cup \{x_0\}$ and its derivative also agrees with the derivative of f at S.*

Lemma 12 *Consider the point $x_1 \in S$. Let f be a real two times differential function and define $y_4(x) = y_2(x) + \frac{f''(x_1)-y_2''(x_1)}{2\pi'^2(x_1)}\pi^2(x)$. Then y_4 is a polynomial with degree $2n$ that agrees with f at S, its derivative also agrees with the derivative of f at S and its second derivative agrees with the second derivative of f at x_1.*

A.3 Derivation of bounds

In this section, we prove one of the two main theoretical results of our paper, namely the bounds on the error of our feature approximations (Theorem 2). We will start with Theorem 13, restating a well known result from Gauss-Hermite quadrature (Hildebrand, 1987). This theorem will inspire the proofs for Theorem 14 and 15, where we prove similar bounds for the more complex cases of first and second order derivatives.

Theorem 13 *Consider the function $cos(\frac{\sqrt{2}}{l}r\omega)$. If we approximate the integral $I(cos(\frac{\sqrt{2}}{l}r\omega) = \int_{-\infty}^{+\infty} e^{-\omega^2} cos(\omega r\frac{\sqrt{2}}{l})d\omega$ by $Q_m(cos(\frac{\sqrt{2}}{l}r\omega)$, the Gauss-Hermite quadrature scheme of order m, we have*

$$|I(cos(\frac{\sqrt{2}}{l}r\omega)) - Q_m(cos(\frac{\sqrt{2}}{l}r\omega))| \leq E_m. \tag{39}$$

Proof 3 *Let $y_{2m}(\omega)$ be a polynomial of degree $2m-1$ that agrees with $cos(\frac{\sqrt{2}}{l}r\omega)$ at S^m and its derivate also agrees with the derivative of $cos(\frac{\sqrt{2}}{l}r\omega)$ at S^m (we know that such a polynomial exists by Lemma 10) . By Lemma 6 we have for every ω that*

$$cos(\frac{\sqrt{2}}{l}r\omega) - y_{2m}(\omega) = (\frac{\sqrt{2}r}{l})^{2m}\frac{cos(\xi)}{(2m)!}h_m^2(\omega) \tag{40}$$

and since $r \leq 1$ and $|cos(\xi)| \leq 1$ we get

$$|cos(\frac{\sqrt{2}}{l}r\omega) - y_{2m}(\omega)| \leq (\frac{\sqrt{2}}{l})^{2m}\frac{h_m^2(\omega)}{(2m)!}. \tag{41}$$

Moreover, $Q_m(cos(\frac{\sqrt{2}}{l}r\omega)) = Q_m(y_{2m}(\omega))$ (because $cos(\frac{\sqrt{2}}{l}r\omega)$ and $y_{2m}(\omega)$ agree at S^m) and $Q_m(y_{2m}(\omega)) = I(y_{2m}(\omega))$ (because y_{2m} has degree less than $2m$) so $Q_m(cos(\frac{\sqrt{2}}{l}r\omega)) = I(y_{2m}(\omega))$ and

$$|I(cos(\frac{\sqrt{2}}{l}r\omega)) - Q_m(cos(\frac{\sqrt{2}}{l}r\omega))| = |I(cos(\frac{\sqrt{2}}{l}r\omega)) - I(y_{2m}(\omega))| \leq \tag{42}$$

$$I(|cos(\omega r\frac{\sqrt{2}}{l}) - y_{2m}(\omega)|) \leq I(\frac{\sqrt{2}}{l})^{2m}\frac{h_m^2(\omega)}{(2m)!}) = \tag{43}$$

$$(\frac{\sqrt{2}}{l})^{2m}\frac{m!\sqrt{\pi}}{2^m(2m!)} \leq E_m. \tag{44}$$

In the preceding proof, the basic idea was to approximate the function with a polynomial (of degree less than $2m$) that agrees with the function at a specific set of points. This gives us a remainder that can be relatively efficiently bounded, leading to a tight error bound. In the following proofs, the main challenge lies in finding the right approximating polynomial (of degree less than $2m$) that yields an easy to handle and efficiently bounded remainder. Once such a polynomial is constructed, we can use the following idea: Let $f(\omega)$ be the function to be approximated by the approximating polynomial $p(\omega)$ and assume that the remainder can be absolutely bounded by the polynomial $s(\omega)$. Then

$$|I(f(\omega)) - Q_m(f(\omega))| = |I(f(\omega)) - I(p(\omega)) + Q_m(p(\omega)) - Q_m(f(\omega))| \leq \tag{45}$$

$$|I(f(\omega)) - I(p(\omega))| + |Q_m(p(\omega)) - Q_m(f(\omega))| \leq \tag{46}$$

$$I(|f(\omega) - p(\omega)|) + Q_m(|f(\omega) - p(\omega)|) \leq I(s(\omega)) + Q_m(s(\omega)). \tag{47}$$

If $s(\omega)$ has degree less than $2m$ so that $I(s(\omega)) = Q_m(s(\omega)))$, then the final bound is $2I(s(\omega))$. Following standard practice, $s(\omega)$ will be chosen as the square of a polynomial. This is motivated by the following observation:

Consider $s(\omega)$ to be the square of a polynomial of degree n. W.l.o.g assume $s(\omega)$ to be monic (has leading coefficient 1). Then $s(\omega) = (h_n(\omega) + q(\omega))^2$ with $deg(q) < n$. Consequently, $I(s(\omega)) = I(h_n^2(\omega)) + I(q^2(\omega)) + 2I(h_n(\omega)q(\omega)) = I(h_n^2(\omega)) + I(q^2(\omega)) \geq I(h_n^2(\omega))$. This means that $s(\omega) = h_n^2(\omega)$ minimizes $I(s(\omega))$ and suggests that in our proofs, the approximating polynomial should agree with the function at the set S^n, so that the remainder $s(\omega)$ is of the form $h_n^2(\omega)$ and gives us good values for $I(s(\omega))$.

We now have all the necessary tools to state and prove Theorem 2. We split the theorem into two parts: Theorem 14 restates the claim of Theorem 2 for the first order derivative, while Theorem 15 restates the claim for the second order derivative.

Theorem 14 *Consider the function $\frac{\sqrt{2}}{l}\omega\sin(\omega r\frac{\sqrt{2}}{l})$. If we approximate the integral $I(\frac{\sqrt{2}}{l}\omega\sin(\omega r\frac{\sqrt{2}}{l})) = \int_{-\infty}^{+\infty} e^{-\omega^2}\frac{\sqrt{2}}{l}\omega\sin(\omega r\frac{\sqrt{2}}{l})d\omega$ by $Q_m(\frac{\sqrt{2}}{l}\omega\sin(\omega r\frac{\sqrt{2}}{l})$, the Gauss-Hermite quadrature scheme of order m, we have*

$$\frac{\sqrt{2}}{l}|I(\omega\sin(\omega r\frac{\sqrt{2}}{l})) - Q_m(\omega\sin(\omega r\frac{\sqrt{2}}{l}))| \leq 8(m-1)E_{m-1}. \tag{48}$$

Proof 4 *Depending on wether m is odd or even we have:*
Suppose m is even. Then $0 \in S^{m-1}$. Let $y_{2m-3}(\omega)$ be a polynomial of degree $2m-4$ that agrees with $sin(\frac{\sqrt{2}}{l}r\omega)$ at S^{m-1} and its derivate also agrees with the derivative of $sin(\frac{\sqrt{2}}{l}r\omega)$ at $S^{m-1}\setminus\{0\}$ (we know that such a polynomial exists by Lemma 11). By Lemma 7 we have for every ω that

$$\omega(sin(\frac{\sqrt{2}}{l}r\omega) - y_{2m-3}(\omega)) = (\frac{\sqrt{2}r}{l})^{2m-3}\frac{cos(\xi)}{(2m-3)!}h_{m-1}^2(\omega) \tag{49}$$

so since $r \leq 1$ and $|cos(\xi)| \leq 1$ we get

$$|\omega sin(\frac{\sqrt{2}}{l}r\omega) - \omega y_{2m-3}(\omega)| \leq (\frac{\sqrt{2}}{l})^{2m-3}\frac{h_{m-1}^2(\omega)}{(2m-3)!}. \tag{50}$$

Consequently, since the weights of a quadrature scheme W_i^m are positive and $Q_m(h_{m-1}^2(\omega)) = I(h_{m-1}^2(\omega))$ (because h_{m-1}^2 has degree less than $2m$), using the above relation we get

$$|Q_m(\omega y_{2m-3}(\omega)) - Q_m(\omega\sin(\omega r\frac{\sqrt{2}}{l})| = |Q_m(\omega y_{2m-3}(\omega) - \omega\sin(\omega r\frac{\sqrt{2}}{l}))| \leq \tag{51}$$

$$Q_m(|\omega y_{2m-3}(\omega) - \omega\sin(\omega r\frac{\sqrt{2}}{l})|) \leq Q_m((\frac{\sqrt{2}}{l})^{2m-3}\frac{h_{m-1}^2(\omega)}{(2m-3)!}) = \tag{52}$$

$$I((\frac{\sqrt{2}}{l})^{2m-3}\frac{h_{m-1}^2(\omega)}{(2m-3)!}) := R \tag{53}$$

where

$$R = (\frac{\sqrt{2}}{l})^{2m-3}\frac{(m-1)!\sqrt{\pi}}{2^{m-1}(2m-3)!}. \tag{54}$$

Similarly,

$$|I(\omega y_{2m-3}(\omega)) - I(\omega\sin(\omega r\frac{\sqrt{2}}{l})| = |I(\omega y_{2m-3}(\omega) - \omega\sin(\omega r\frac{\sqrt{2}}{l}))| \leq \tag{55}$$

$$I(|\omega y_{2m-3}(\omega) - \omega\sin(\omega r\frac{\sqrt{2}}{l})|) \leq I((\frac{\sqrt{2}}{l})^{2m-3}\frac{h_{m-1}^2(\omega)}{(2m-3)!}) = R. \tag{56}$$

Finally, using that $Q_m(\omega y_{2m-3}(\omega)) = I(\omega y_{2m-3}(\omega))$ (because $\omega y_{2m-3}(\omega)$ has degree less than $2m$) we get

$$\frac{\sqrt{2}}{l}|I(\omega\sin(\omega r\frac{\sqrt{2}}{l})) - Q_m(\omega\sin(\omega r\frac{\sqrt{2}}{l}))| = \tag{57}$$

$$\frac{\sqrt{2}}{l}|I(\omega\sin(\omega r\frac{\sqrt{2}}{l})) - I(\omega y_{2m-3}(\omega)) + Q_m(\omega y_{2m-3}(\omega)) - Q_m(\omega\sin(\omega r\frac{\sqrt{2}}{l}))| \leq \tag{58}$$

$$\frac{\sqrt{2}}{l}(|I(\omega\sin(\omega r\frac{\sqrt{2}}{l})) - I(\omega y_{2m-3}(\omega))| + |Q_m(\omega y_{2m-3}(\omega)) - Q_m(\omega\sin(\omega r\frac{\sqrt{2}}{l}))|) \leq \tag{59}$$

$$\frac{\sqrt{2}}{l}2R = 2(\frac{\sqrt{2}}{l})^{2m-2}\frac{(m-1)!\sqrt{\pi}}{2^{m-1}(2m-3)!} \leq 4(m-1)E_{m-1}. \tag{60}$$

Suppose m is odd. Then $0 \in S^{m-2}$. Let $y_{2m-3}(\omega)$ be a polynomial of degree $2m-4$ that agrees with $sin(\frac{\sqrt{2}}{l}r\omega)$ at S^{m-2} and its derivate also agrees with the derivative of $sin(\frac{\sqrt{2}}{l}r\omega)$ at S^{m-2} and its second derivative agrees at 0 with with the second derivative of $sin(\frac{\sqrt{2}}{l}r\omega)$ (we know that such a polynomial exists by Lemma 12). By Lemma 8 we have for every ω that

$$\omega(sin(\frac{\sqrt{2}}{l}r\omega) - y_{2m-3}(\omega)) = (\frac{\sqrt{2}r}{l})^{2m-3}\frac{cos(\xi)}{(2m-3)!}\omega^2 h_{m-2}^2(\omega) \tag{61}$$

so since $r \leq 1$ and $|cos(\xi)| \leq 1$ we get

$$|\omega\sin(\frac{\sqrt{2}}{l}r\omega) - \omega y_{2m-3}(\omega)| \leq (\frac{\sqrt{2}}{l})^{2m-3}\frac{\omega^2 h_{m-2}^2(\omega)}{(2m-3)!}. \tag{62}$$

Consequently, since the weights of a quadrature scheme W_i^m are positive and $Q_m(\omega^2 h_{m-2}^2(\omega)) = I(\omega^2 h_{m-2}^2(\omega))$ (because $\omega^2 h_{m-2}^2$ has degree less than $2m$), using the above relation we get

$$|Q_m(\omega y_{2m-3}(\omega)) - Q_m(\omega\sin(\omega r\frac{\sqrt{2}}{l})| = |Q_m(\omega y_{2m-3}(\omega) - \omega\sin(\omega r\frac{\sqrt{2}}{l}))| \leq \tag{63}$$

$$Q_m(|\omega y_{2m-3}(\omega) - \omega\sin(\omega r\frac{\sqrt{2}}{l})|) \leq Q_m((\frac{\sqrt{2}}{l})^{2m-3}\frac{\omega^2 h_{m-2}^2(\omega)}{(2m-3)!}) = \tag{64}$$

$$I((\frac{\sqrt{2}}{l})^{2m-3}\frac{\omega^2 h_{m-2}^2(\omega)}{(2m-3)!}) := R \tag{65}$$

where

$$R = (\frac{\sqrt{2}}{l})^{2m-3}\frac{I(\omega^2 h_{m-2}^2(\omega))}{(2m-3)!}.$$

We will now use the identity

$$xh_n(x) = h_{n+1}(x) + \frac{n}{2}h_{n-1}(x)$$

and that $I(h_{n+1}(\omega)h_{n-1}(\omega)) = 0$ (normality) so by squaring we have

$$I(\omega^2 h_{m-2}^2(\omega)) = I(h_{m-1}^2(\omega)) + (\frac{m-2}{2})^2 I(h_{m-3}^2(\omega)) \leq 2\frac{(m-1)!\sqrt{\pi}}{2^{m-1}}. \tag{66}$$

So

$$R \leq 2(\frac{\sqrt{2}}{l})^{2m-3}\frac{(m-1)!\sqrt{\pi}}{2^{m-1}(2m-3)!}. \tag{67}$$

Similarly,

$$|I(\omega y_{2m-3}(\omega)) - I(\omega\sin(\omega r\frac{\sqrt{2}}{l})| = |I(\omega y_{2m-3}(\omega) - \omega\sin(\omega r\frac{\sqrt{2}}{l}))| \leq \tag{68}$$

$$I(|\omega y_{2m-3}(\omega) - \omega\sin(\omega r\frac{\sqrt{2}}{l})|) \leq I((\frac{\sqrt{2}}{l})^{2m-3}\frac{\omega^2 h_{m-2}^2(\omega)}{(2m-3)!}) = R. \tag{69}$$

Finally, using that $Q_m(\omega y_{2m-3}(\omega)) = I(\omega y_{2m-3}(\omega))$ (because $\omega y_{2m-3}(\omega)$ has degree less than $2m$) we get

$$\frac{\sqrt{2}}{l}|I(\omega\sin(\omega r\frac{\sqrt{2}}{l})) - Q_m(\omega\sin(\omega r\frac{\sqrt{2}}{l}))| = \tag{70}$$

$$\frac{\sqrt{2}}{l}|I(\omega\sin(\omega r\frac{\sqrt{2}}{l})) - I(\omega y_{2m-3}(\omega)) + Q_m(\omega y_{2m-3}(\omega)) - Q_m(\omega\sin(\omega r\frac{\sqrt{2}}{l}))| \leq \tag{71}$$

$$\frac{\sqrt{2}}{l}(|I(\omega\sin(\omega r\frac{\sqrt{2}}{l})) - I(\omega y_{2m-3}(\omega))| + |Q_m(\omega y_{2m-3}(\omega)) - Q_m(\omega\sin(\omega r\frac{\sqrt{2}}{l}))|) \leq \tag{72}$$

$$\frac{\sqrt{2}}{l}2R = 4(\frac{\sqrt{2}}{l})^{2m-2}\frac{(m-1)!\sqrt{\pi}}{2^{m-1}(2m-3)!} \leq 8(m-1)E_{m-1}. \tag{73}$$

Theorem 15 *Consider the function $\frac{2}{l^2}\omega^2\cos(\omega r\frac{\sqrt{2}}{l})$. If we approximate the integral $I(\frac{2}{l^2}\omega^2\cos(\omega r\frac{\sqrt{2}}{l})) = \int_{-\infty}^{+\infty} e^{-\omega^2}\frac{2}{l^2}\omega^2\cos(\omega r\frac{\sqrt{2}}{l})d\omega$ by $Q_m(\frac{2}{l^2}\omega^2\cos(\omega r\frac{\sqrt{2}}{l}))$, the Gauss-Hermite quadrature scheme of order m, we have*

$$\frac{2}{l^2}|I(\omega^2\cos(\omega r\frac{\sqrt{2}}{l}) - Q_m(\omega^2\cos(\omega r\frac{\sqrt{2}}{l})| \leq \frac{4}{l^2}(m-1)E_{m-2}. \tag{74}$$

Proof 5 *Depending on wether m is odd or even we have:*
Suppose m is even. Then $0 \in S^{m-1}$. Let $y_{2m-4}(\omega)$ be a polynomial of degree $2m-5$ that agrees with $\cos(\frac{\sqrt{2}}{l}r\omega)$ at $S^{m-1}\setminus\{0\}$ and its derivate also agrees with the derivative of $\cos(\frac{\sqrt{2}}{l}r\omega)$ at $S^{m-1}\setminus\{0\}$ (we know that such a polynomial exists by Lemma 11). By Lemma 7 we have for every ω that

$$\omega^2(\cos(\frac{\sqrt{2}}{l}r\omega) - y_{2m-4}(\omega)) = (\frac{\sqrt{2}r}{l})^{2m-4}\frac{\cos(\xi)}{(2m-4)!}h_{m-1}^2(\omega) \tag{75}$$

so since $r \leq 1$ and $|cos(\xi)| \leq 1$ we get

$$|\omega^2\cos(\frac{\sqrt{2}}{l}r\omega) - \omega^2 y_{2m-4}(\omega)| \leq (\frac{\sqrt{2}}{l})^{2m-4}\frac{h_{m-1}^2(\omega)}{(2m-4)!}. \tag{76}$$

Consequently, since the weights of a quadrature scheme W_i^m are positive and $Q_m(h_{m-1}^2(\omega)) = I(h_{m-1}^2(\omega))$ (because h_{m-1}^2 has degree less than $2m$), using the above relation we get

$$|Q_m(\omega^2 y_{2m-4}(\omega)) - Q_m(\omega^2 \cos(\omega r\frac{\sqrt{2}}{l})| = |Q_m(\omega^2 y_{2m-4}(\omega) - \omega^2 \cos(\omega r\frac{\sqrt{2}}{l}))| \leq \tag{77}$$

$$Q_m(|\omega^2 y_{2m-4}(\omega) - \omega^2 \cos(\omega r\frac{\sqrt{2}}{l})|) \leq Q_m((\frac{\sqrt{2}}{l})^{2m-4}\frac{h_{m-1}^2(\omega)}{(2m-4)!}) = \tag{78}$$

$$I((\frac{\sqrt{2}}{l})^{2m-4}\frac{h_{m-1}^2(\omega)}{(2m-4)!}) := R \tag{79}$$

where

$$R = (\frac{\sqrt{2}}{l})^{2m-4}\frac{(m-1)!\sqrt{\pi}}{2^{m-1}(2m-4)!}. \tag{80}$$

Similarly,

$$|I(\omega^2 y_{2m-4}(\omega)) - I(\omega^2 \cos(\omega r\frac{\sqrt{2}}{l})| = |I(\omega^2 y_{2m-4}(\omega) - \omega^2 \cos(\omega r\frac{\sqrt{2}}{l}))| \leq \tag{81}$$

$$I(|\omega^2 y_{2m-4}(\omega) - \omega^2 \cos(\omega r\frac{\sqrt{2}}{l})|) \leq I((\frac{\sqrt{2}}{l})^{2m-4}\frac{h_{m-1}^2(\omega)}{(2m-4)!}) = R. \tag{82}$$

Finally using that $Q_m(\omega^2 y_{2m-4}(\omega)) = I(\omega^2 y_{2m-4}(\omega))$ (because $\omega^2 y_{2m-4}(\omega)$ has degree less than $2m$) and following exactly the same procedure as in the previous proof we get

$$\frac{2}{l^2}|I(\omega^2 \cos(\omega r\frac{\sqrt{2}}{l}) - Q_m(\omega^2 \cos(\omega r\frac{\sqrt{2}}{l})| \leq \frac{2}{l^2}2R = 2(\frac{\sqrt{2}}{l})^{2m-2}\frac{(m-1)!\sqrt{\pi}}{2^{m-1}(2m-4)!} \leq \frac{2}{l^2}(m-1)E_{m-2}. \tag{83}$$

Suppose m is odd. Let $y_{2m-4}(\omega)$ be a polynomial of degree $2m-5$ that agrees with $\cos(\frac{\sqrt{2}}{l}r\omega)$ at S^{m-2} and its derivate also agrees with the derivative of $\cos(\frac{\sqrt{2}}{l}r\omega)$ at S^{m-2} (we know that such a polynomial exists by Lemma 10). By Lemma 6 we have for every ω that

$$\omega^2(\cos(\frac{\sqrt{2}}{l}r\omega) - y_{2m-4}(\omega)) = (\frac{\sqrt{2}r}{l})^{2m-4}\frac{cos(\xi)}{(2m-4)!}\omega^2 h_{m-2}^2(\omega) \tag{84}$$

so since $r \leq 1$ and $|cos(\xi)| \leq 1$ we get

$$|\omega^2 \cos(\frac{\sqrt{2}}{l}r\omega) - \omega^2 y_{2m-4}(\omega)| \leq (\frac{\sqrt{2}}{l})^{2m-4}\frac{\omega^2 h_{m-2}^2(\omega)}{(2m-4)!}. \tag{85}$$

Consequently, since the weights of a quadrature scheme W_i^m are positive and $Q_m(\omega^2 h_{m-2}^2(\omega)) = I(\omega^2 h_{m-2}^2(\omega))$ (because $\omega^2 h_{m-2}^2$ has degree less than $2m$), using the above relation we get

$$|Q_m(\omega^2 y_{2m-4}(\omega)) - Q_m(\omega^2 \cos(\omega r\frac{\sqrt{2}}{l})| = |Q_m(\omega^2 y_{2m-4}(\omega) - \omega^2 \cos(\omega r\frac{\sqrt{2}}{l}))| \leq \tag{86}$$

$$Q_m(|\omega^2 y_{2m-4}(\omega) - \omega^2 \cos(\omega r\frac{\sqrt{2}}{l})|) \leq Q_m((\frac{\sqrt{2}}{l})^{2m-4}\frac{\omega^2 h_{m-2}^2(\omega)}{(2m-4)!}) = \tag{87}$$

$$I((\frac{\sqrt{2}}{l})^{2m-4}\frac{\omega^2 h_{m-2}^2(\omega)}{(2m-4)!}) := R \tag{88}$$

where

$$R = (\frac{\sqrt{2}}{l})^{2m-4}\frac{I(\omega^2 h_{m-2}^2(\omega))}{(2m-4)!}.$$

We will now use the identity

$$xh_n(x) = h_{n+1}(x) + \frac{n}{2}h_{n-1}(x)$$

and that $I(h_{n+1}(\omega)h_{n-1}(\omega)) = 0$ *(normality) so by squaring we have*

$$I(\omega^2 h_{m-2}^2(\omega)) = I(h_{m-1}^2(\omega)) + (\frac{m-2}{2})^2 I(h_{m-3}^2(\omega)) \leq 2\frac{(m-1)!\sqrt{\pi}}{2^{m-1}}. \tag{89}$$

So

$$R \leq 2(\frac{\sqrt{2}}{l})^{2m-4}\frac{(m-1)!\sqrt{\pi}}{2^{m-1}(2m-4)!}. \tag{90}$$

Similarly,

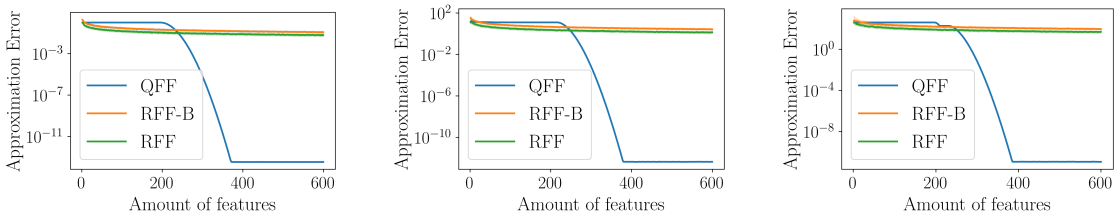
$$|I(\omega^2 y_{2m-4}(\omega)) - I(\omega^2\cos(\omega r\frac{\sqrt{2}}{l})| = |I(\omega^2 y_{2m-4}(\omega) - \omega^2\cos(\omega r\frac{\sqrt{2}}{l}))| \leq \tag{91}$$

$$I(|\omega^2 y_{2m-4}(\omega) - \omega^2\cos(\omega r\frac{\sqrt{2}}{l})|) \leq I((\frac{\sqrt{2}}{l})^{2m-4}\frac{\omega^2 h_{m-2}^2(\omega)}{(2m-4)!}) = R. \tag{92}$$

Finally, using that $Q_m(\omega^2 y_{2m-4}(\omega)) = I(\omega^2 y_{2m-4}(\omega))$ *(because* $\omega^2 y_{2m-4}(\omega)$ *has degree less than* $2m$*) and following exactly the same procedure as in the previous proof we get*

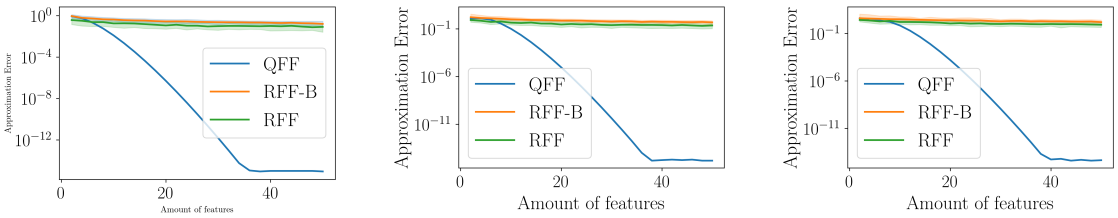
$$\frac{2}{l^2}|I(\omega^2\cos(\omega r\frac{\sqrt{2}}{l}) - Q_m(\omega^2\cos(\omega r\frac{\sqrt{2}}{l})| \leq \frac{2}{l^2}2R = 4(\frac{\sqrt{2}}{l})^{2m-2}\frac{(m-1)!\sqrt{\pi}}{2^{m-1}(2m-4)!} \leq \frac{4}{l^2}(m-1)E_{m-2}. \tag{93}$$

Appendix B. Kernel Approximation Additional Plots



(a) $k(r)$ (b) $k'(r)$ (c) $k''(r)$

Figure 6: Comparing the maximum error of different feature expansions over $r \in [0, 1]$. For the random feature expansions, we show median as well as 12.5% and 87.5% quantiles over 100 random samples. Due to the exponential decay of the error of the QFF approximation, this stochasticity is barely visible. As given by the theoretical analysis, the error is a bit higher for the derivatives, but still decaying exponentially. In this plot, we set $l = 0.05$.



(a) $k(r)$ (b) $k'(r)$ (c) $k''(r)$

Figure 7: Comparing the maximum error of different feature expansions over $r \in [0, 1]$. For the random feature expansions, we show median as well as 12.5% and 87.5% quantiles over 100 random samples. Due to the exponential decay of the error of the QFF approximation, this stochasticity is barely visible. As given by the theoretical analysis, the error is a bit higher for the derivatives, but still decaying exponentially. In this plot, we set $l = 0.5$.

Appendix C. GP Regression with Derivatives

In this section, we will introduce the necessary notation and the proof for our theoretical results on GP regression with derivatives.

C.1 Notation

Consider the problem of Gaussian Process regression, using zero mean prior and the RBF kernel function $k_{\boldsymbol{\phi}}(x,y) := \rho e^{-\frac{(x-y)^2}{2l^2}}$ for some fixed hyperparameters $\boldsymbol{\phi} = (\rho, l)$, which denote the variance and the lengthscale. Suppose we are given at n observation points $\boldsymbol{t} := \boldsymbol{t}_n = (t_1, \ldots, t_n)$ the n dimensional (column) vectors $\boldsymbol{y} := \boldsymbol{x} + \boldsymbol{\epsilon}_{\sigma^2}$ and $\boldsymbol{F} := \dot{\boldsymbol{x}} + \boldsymbol{\epsilon}_{\gamma}$ of noisy state and noisy state derivative observations. Our goal is to infer at the observation point T the values of state $x := x(T)$ and state derivative $\dot{x} := \dot{x}(T)$.

We repeat some of the definitions already given in this paper and this appendix as well as add a few new ones relevant for this section only:

Let $\boldsymbol{k}_{\boldsymbol{\phi}}(\boldsymbol{t}, T)$ denote the n dimensional kernel (column) vector, i.e.

$$\boldsymbol{k}_{\boldsymbol{\phi}}(\boldsymbol{t}, T)_i := k_{\boldsymbol{\phi}}(t_i, T). \tag{94}$$

Let ${}'k_{\boldsymbol{\phi}}(x,y)$ denote the partial derivative of $k_{\boldsymbol{\phi}}$ w.r.t. its first argument, i.e.

$${}'k_{\boldsymbol{\phi}}(x,y) := \frac{\partial}{\partial a} k_{\boldsymbol{\phi}}(a,b)|_{a=x,b=y}. \tag{95}$$

Let ${}'\boldsymbol{k}_{\boldsymbol{\phi}}(\boldsymbol{t}, T)$ denote the n dimensional kernel derivative (column) vector, i.e.

$${}'\boldsymbol{k}_{\boldsymbol{\phi}}(\boldsymbol{t}, T)_i := {}'k_{\boldsymbol{\phi}}(t_i, T). \tag{96}$$

Let $k'_{\boldsymbol{\phi}}(x,y)$ denote the partial derivative of $k_{\boldsymbol{\phi}}$ w.r.t. its second argument, i.e.

$$k'_{\boldsymbol{\phi}}(x,y) := \frac{\partial}{\partial b} k_{\boldsymbol{\phi}}(a,b)|_{a=x,b=y}. \tag{97}$$

Let $\boldsymbol{k}'_{\boldsymbol{\phi}}(\boldsymbol{t}, T)$ denote the n dimensional kernel derivative (column) vector, i.e.

$$\boldsymbol{k}'_{\boldsymbol{\phi}}(\boldsymbol{t}, T)_i := k'_{\boldsymbol{\phi}}(t_i, T). \tag{98}$$

Let $k''_{\boldsymbol{\phi}}(x,y)$ denote the mixed partial derivative of $k_{\boldsymbol{\phi}}$, i.e.

$$k''_{\boldsymbol{\phi}}(x,y) := \frac{\partial^2}{\partial a \partial b} k_{\boldsymbol{\phi}}(a,b)|_{a=x,b=y}. \tag{99}$$

Let $\boldsymbol{k}''_{\boldsymbol{\phi}}(\boldsymbol{t}, T)$ denote the n dimensional kernel derivative (column) vector, i.e.

$$\boldsymbol{k}''_{\boldsymbol{\phi}}(\boldsymbol{t}, T)_i := k''_{\boldsymbol{\phi}}(t_i, T). \tag{100}$$

Let $\boldsymbol{C}_{\boldsymbol{\phi}}$ denote the $n \times n$ covariance kernel matrix, whose elements are given by

$$[\boldsymbol{C}_{\boldsymbol{\phi}}]_{i,j} := k_{\boldsymbol{\phi}}(t_i, t_j). \tag{101}$$

Let ${}'\boldsymbol{C}_{\boldsymbol{\phi}}$ denote the kernel derivative matrix, whose elements are given by

$$\left[{}'\boldsymbol{C}_{\boldsymbol{\phi}}\right]_{i,j} := \frac{\partial}{\partial a} k_{\boldsymbol{\phi}}(a,b)|_{a=t_i,b=t_j}. \tag{102}$$

Let $\boldsymbol{C}'_{\boldsymbol{\phi}}$ denote the kernel derivative matrix, whose elements are given by

$$\left[\boldsymbol{C}'_{\boldsymbol{\phi}}\right]_{i,j} := \frac{\partial}{\partial b} k_{\boldsymbol{\phi}}(a,b)|_{a=t_i,b=t_j}. \tag{103}$$

Let $\boldsymbol{C}''_{\boldsymbol{\phi}}$ denote the mixed kernel derivative matrix, whose elements are given by

$$\left[\boldsymbol{C}''_{\boldsymbol{\phi}}\right]_{i,j} := \frac{\partial^2}{\partial a \partial b} k_{\boldsymbol{\phi}}(a,b)|_{a=t_i,b=t_j}. \tag{104}$$

Let $\hat{\boldsymbol{K}}_{\boldsymbol{\phi}}$ denote the sum of the $2n \times 2n$ block matrix with the covariance matrix and its derivatives plus the diagonal noise matrix, i.e.

$$\hat{\boldsymbol{K}}_{\boldsymbol{\phi}} := \begin{pmatrix} \boldsymbol{C}_{\boldsymbol{\phi}} & \boldsymbol{C}'_{\boldsymbol{\phi}} \\ {}'\boldsymbol{C}_{\boldsymbol{\phi}} & \boldsymbol{C}''_{\boldsymbol{\phi}} \end{pmatrix} + \begin{pmatrix} \sigma^2 \mathbb{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \gamma \mathbb{I}_n \end{pmatrix}. \tag{105}$$

Let $\hat{\boldsymbol{k}}_{\phi}(\boldsymbol{t}, T)$ denote the $2n$ dimensional (column) vector, which is a concatenation of $\boldsymbol{k}_{\boldsymbol{\phi}}(\boldsymbol{t}, T)$ and ${}'\boldsymbol{k}_{\boldsymbol{\phi}}(\boldsymbol{t}, T)$, i.e.

$$\hat{\boldsymbol{k}}_{\phi}(\boldsymbol{t}, T) := \begin{pmatrix} \boldsymbol{k}_{\boldsymbol{\phi}}(\boldsymbol{t}, T) \\ {}'\boldsymbol{k}_{\boldsymbol{\phi}}(\boldsymbol{t}, T) \end{pmatrix}. \tag{106}$$

Let $\hat{\boldsymbol{k}}'_{\phi}(\boldsymbol{t}, T)$ denote the $2n$ dimensional (column) vector, which is a concatenation of $\boldsymbol{k}'_{\boldsymbol{\phi}}(\boldsymbol{t}, T)$ and $\boldsymbol{k}''_{\boldsymbol{\phi}}(\boldsymbol{t}, T)$, i.e.

$$\hat{\boldsymbol{k}}'_{\phi}(\boldsymbol{t}, T) := \begin{pmatrix} \boldsymbol{k}'_{\boldsymbol{\phi}}(\boldsymbol{t}, T) \\ \boldsymbol{k}''_{\boldsymbol{\phi}}(\boldsymbol{t}, T) \end{pmatrix}. \tag{107}$$

Finally, we are able to write down the formulas for the scalar predictive mean and covariance at a new point T. Here, we let μ denote the mean of the state, μ' denote the mean of the derivative, Σ denote the variance of the state and Σ' the variance of the derivative prediction. They are given by

$$\mu(T) = \hat{\boldsymbol{k}}_{\phi}(\boldsymbol{t}, T)^T \hat{\boldsymbol{K}}_{\boldsymbol{\phi}}^{-1} \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{F} \end{pmatrix}, \tag{108}$$

$$\mu'(T) = \hat{\boldsymbol{k}}'_{\phi}(\boldsymbol{t}, T)^T \hat{\boldsymbol{K}}_{\boldsymbol{\phi}}^{-1} \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{F} \end{pmatrix}, \tag{109}$$

$$\Sigma(T) = k(T,T) - \hat{\boldsymbol{k}}_{\phi}(\boldsymbol{t}, T)^T \hat{\boldsymbol{K}}_{\boldsymbol{\phi}}^{-1} \hat{\boldsymbol{k}}_{\phi}(\boldsymbol{t}, T), \tag{110}$$

$$\Sigma'(T) = k''(T,T) - \hat{\boldsymbol{k}}'_{\phi}(\boldsymbol{t}, T)^T \hat{\boldsymbol{K}}_{\boldsymbol{\phi}}^{-1} \hat{\boldsymbol{k}}'_{\phi}(\boldsymbol{t}, T). \tag{111}$$

C.2 Proof of Theorem 3

In this Section, we will prove the main theorem regarding the approximation error of GP regression with derivatives. For ease of reference, we will restate it as Theorem 16. Throughout this section, we will use $\tilde{a}$ to denote the feature approximation of any scalar, vector or matrix a.

Theorem 16 *Let us consider an RBF kernel with hyperparameters (ρ, l) and domain $[0,1]$. Define $e_{\tilde{\mu}}$, $e_{\tilde{\Sigma}}$, $e_{\tilde{\mu}'}$ and $e_{\tilde{\Sigma}'}$ as the absolute error between the feature approximations and the corresponding accurate quantities of the means and covariances of Equations (4) and (5). For each $\tau \in [0,1]$, define e_{tot} as the maximum of these four errors. Define $c \coloneqq \min(\gamma, \sigma^2)$ and $R \coloneqq \max(||\boldsymbol{y}||_\infty, ||\boldsymbol{F}||_\infty)$. Let $C > 0$. Let us consider a QFF approximation scheme of order $m \geq 3 + \max\left(\frac{e}{2l^2}, \log\left(\frac{270n^2\rho^3R}{l^8c^2C}\right)\right)$. Then, it holds for all $\tau \in [0,1]$ that $e_{tot} \leq C$.*

Proof 6 *Suppose that we apply a QFF approximation scheme of order m for the functions $k_{\boldsymbol{\phi}}, {}'k_{\boldsymbol{\phi}}, k'_{\boldsymbol{\phi}}, k''_{\boldsymbol{\phi}}$, which gives us a deterministic and uniform (over their domain) approximation guarantee of absolute error less than $\epsilon \coloneqq \epsilon_\phi(m)$ (for any of them). W.l.o.g we can assume that the domain of these functions is $[0,1]^2$, $l \leq 1$ and $\rho \geq 1$ (so also $0 \leq T, t_1, \dots t_n \leq 1$). Moreover, we assume that $|\boldsymbol{y}|_{max}, |\boldsymbol{F}|_{max} \leq R$, for some positive constant R.*

Let $\boldsymbol{E}_1$ be the error (matrix) when approximating $\hat{\boldsymbol{K}}_{\boldsymbol{\phi}}$, i.e.

$$\boldsymbol{E}_1 \coloneqq \hat{\boldsymbol{K}}_{\boldsymbol{\phi}} - \tilde{\hat{\boldsymbol{K}}}_{\boldsymbol{\phi}}. \tag{112}$$

Let $\boldsymbol{E}_2$be the error (matrix) when approximating $\hat{\boldsymbol{K}}_{\boldsymbol{\phi}}^{-1}$, i.e.

$$\boldsymbol{E}_2 \coloneqq \hat{\boldsymbol{K}}_{\boldsymbol{\phi}}^{-1} - \tilde{\hat{\boldsymbol{K}}}_{\boldsymbol{\phi}^{-1}}. \tag{113}$$

Let $\boldsymbol{e}_1$ be the error (vector) when approximating $\hat{\boldsymbol{k}}_\phi(\boldsymbol{t}, T)$, i.e.

$$\boldsymbol{e}_1 \coloneqq \hat{\boldsymbol{k}}_\phi(\boldsymbol{t}, T) - \tilde{\hat{\boldsymbol{k}}}_\phi(\boldsymbol{t}, T). \tag{114}$$

Let $\boldsymbol{e}_2$ be the error (vector) when approximating $\hat{\boldsymbol{k}}_\phi(\boldsymbol{t}, T)$, i.e.

$$\boldsymbol{e}_2 \coloneqq \hat{\boldsymbol{k}}'_\phi(\boldsymbol{t}, T) - \tilde{\hat{\boldsymbol{k}}}'_{\boldsymbol{\phi}}(\boldsymbol{t}, T). \tag{115}$$

Given these assumptions and definitions, we can introduce the following bounds:

$$||\boldsymbol{e}_1|| \leq \sqrt{2n}\epsilon, \tag{116}$$

$$||\boldsymbol{e}_2|| \leq \sqrt{2n}\epsilon, \tag{117}$$

$$\sigma_1(\boldsymbol{E}_1) \leq 2n\epsilon, \tag{118}$$

$$\left|\left|\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{F} \end{pmatrix}\right|\right| \leq \sqrt{2n}R. \tag{119}$$

Using the fact that $k_{\boldsymbol{\phi}} \leq \rho$, ${}'k_{\boldsymbol{\phi}} \leq \frac{\rho}{l}$ and $k''_{\boldsymbol{\phi}} \leq \frac{2\rho}{l^2}$, we can also get

$$\left|\left|\hat{\boldsymbol{k}}_\phi(\boldsymbol{t}, T)\right|\right| \leq \sqrt{2n}\frac{\rho}{l}, \tag{120}$$

$$\left|\left|\hat{\boldsymbol{k}}'_\phi(\boldsymbol{t}, T)\right|\right| \leq \sqrt{2n}\frac{2\rho}{l^2}. \tag{121}$$

We know that the matrix $\begin{pmatrix} \boldsymbol{C}_{\boldsymbol{\phi}} & \boldsymbol{C}'_{\boldsymbol{\phi}} \\ {}'\boldsymbol{C}_{\boldsymbol{\phi}} & \boldsymbol{C}''_{\boldsymbol{\phi}} \end{pmatrix}$ and its QFF approximation are symmetric positive semi definite and thus, their smallest singular value is non negative. Also, $\begin{pmatrix} \sigma^2\mathbb{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \gamma\mathbb{I}_n \end{pmatrix}$ has smallest

singular value c. Thus, both $\hat{\boldsymbol{K}}$ and $\tilde{\hat{\boldsymbol{K}}}$ are symmetric, positive definite matrices and we have

$$\sigma_1(\hat{\boldsymbol{K}}_{\boldsymbol{\phi}}^{-1}) \leq \frac{1}{c}, \tag{122}$$

$$\sigma_1(\tilde{\hat{\boldsymbol{K}}}_{\boldsymbol{\phi}}^{-1}) \leq \frac{1}{c}. \tag{123}$$

Finally, we can use the Woodbury identity to get $\boldsymbol{E}_2 = -\hat{\boldsymbol{K}}_{\boldsymbol{\phi}}^{-1}\boldsymbol{E}_1\tilde{\hat{\boldsymbol{K}}}_{\boldsymbol{\phi}^{-1}}$, justifying

$$\sigma_1(\boldsymbol{E}_2) \leq \frac{2n}{c^2}\epsilon. \tag{124}$$

We now have all the building blocks needed to provide a bound for $e_{\tilde{\mu}}$: Using all the above notations, assumptions and results we get

$$|\epsilon_\mu| = \left|\hat{\boldsymbol{k}}_\phi(\boldsymbol{t},T)^T\hat{\boldsymbol{K}}_{\boldsymbol{\phi}}^{-1}\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{F}\end{pmatrix} - (\hat{\boldsymbol{k}}_\phi(\boldsymbol{t},T)-\boldsymbol{e}_1)^T(\hat{\boldsymbol{K}}_{\boldsymbol{\phi}}^{-1}-\boldsymbol{E}_1)\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{F}\end{pmatrix}\right| \tag{125}$$

$$= \left|(\boldsymbol{e}_1^T\hat{\boldsymbol{K}}_{\boldsymbol{\phi}}^{-1} + \hat{\boldsymbol{k}}_\phi(\boldsymbol{t},T)^T\boldsymbol{E}_1 - \boldsymbol{e}_1^T\boldsymbol{E}_1)\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{F}\end{pmatrix}\right| \tag{126}$$

$$\leq \sqrt{2n}R\left(\frac{\sqrt{2n}}{c}\epsilon + \frac{\rho}{l}\sqrt{2n}2n\epsilon + \sqrt{2n}2n\epsilon^2\right) \tag{127}$$

$$\leq 10\frac{n^2R\rho}{lc}\epsilon. \tag{128}$$

Similarly, we get

$$|\epsilon_{\mu'}| = \left|(\boldsymbol{e}_2^T\hat{\boldsymbol{K}}_{\boldsymbol{\phi}}^{-1} + \hat{\boldsymbol{k}}'_\phi(\boldsymbol{t},T)^T\boldsymbol{E}_1 - \boldsymbol{e}_2^T\boldsymbol{E}_1)\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{F}\end{pmatrix}\right| \tag{129}$$

$$\leq \sqrt{2n}R\left(\frac{\sqrt{2n}}{c}\epsilon + \frac{\rho}{l^2}\sqrt{2n}4n\epsilon + \sqrt{2n}2n\epsilon^2\right) \tag{130}$$

$$\leq 14\frac{n^2R\rho}{l^2c}\epsilon. \tag{131}$$

Moreover,

$$|\epsilon_\Sigma| = |k(T,T) - \hat{\boldsymbol{k}}_\phi(\boldsymbol{t},T)^T\hat{\boldsymbol{K}}_{\boldsymbol{\phi}}^{-1}\hat{\boldsymbol{k}}_\phi(\boldsymbol{t},T) - \tilde{k}(T,T) \tag{132}$$

$$+ (\hat{\boldsymbol{k}}_\phi(\boldsymbol{t},T)-\boldsymbol{e}_1)^T(\hat{\boldsymbol{K}}_{\boldsymbol{\phi}}^{-1}-\boldsymbol{E}_1)(\hat{\boldsymbol{k}}_\phi(\boldsymbol{t},T)-\boldsymbol{e}_1)| \tag{133}$$

$$\leq \epsilon + |-2\boldsymbol{e}_1^T\tilde{\hat{\boldsymbol{K}}}_{\boldsymbol{\phi}^{-1}}\hat{\boldsymbol{k}}_\phi(\boldsymbol{t},T) + \boldsymbol{e}_1^T\tilde{\hat{\boldsymbol{K}}}_{\boldsymbol{\phi}}^{-1}\boldsymbol{e}_1 - \hat{\boldsymbol{k}}_\phi(\boldsymbol{t},T)^T\boldsymbol{E}_2\hat{\boldsymbol{k}}_\phi(\boldsymbol{t},T)| \tag{134}$$

$$\leq \epsilon + \frac{4n\rho}{lc}\epsilon + \frac{2n}{c}\epsilon^2 + \frac{4n^2\rho^2}{l^2c^2}\epsilon \tag{135}$$

$$\leq \frac{14n^2\rho^2}{l^2c^2}\epsilon \tag{136}$$

and

$$|\epsilon_{\Sigma'}| \leq \epsilon + |-2\boldsymbol{e}_2^T \tilde{\hat{\boldsymbol{K}}}_{\boldsymbol{\phi}}^{-1} \hat{\boldsymbol{k}}'_{\phi}(\boldsymbol{t}, T) + \boldsymbol{e}_2^T \tilde{\hat{\boldsymbol{K}}}_{\boldsymbol{\phi}}^{-1} \boldsymbol{e}_2 - \hat{\boldsymbol{k}}'_{\phi}(\boldsymbol{t}, T)^T \boldsymbol{E}_2 \hat{\boldsymbol{k}}'_{\phi}(\boldsymbol{t}, T)| \tag{137}$$

$$\leq \epsilon + \frac{8n\rho}{l^2 c}\epsilon + \frac{2n}{c}\epsilon^2 + \frac{16n^2\rho^2}{l^4 c^2}\epsilon \tag{138}$$

$$\leq \frac{27n^2\rho^2}{l^4 c^2}\epsilon. \tag{139}$$

To summarize all of these upper bounds, we can observe that

$$|e_{tot}| \leq \frac{27n^2\rho^2 R}{l^4 c^2}\epsilon. \tag{140}$$

Let us now fix a constant $0 < C < 1$ *and define* $M \coloneqq m - 3$*. Let us choose M such that*

$$M \geq \max(\frac{e}{2l^2}, \log(\frac{270n^2\rho^3 R}{l^8 c^2 C})). \tag{141}$$

With this choice, it holds that

$$(\frac{e}{4l^2 M})^M \leq \frac{l^8 c^2}{270n^2\rho^3 R} C \tag{142}$$

$$\implies \frac{2e\sqrt{\pi}\rho}{l^4}(\frac{e}{4l^2 M})^M \leq \frac{l^4 c^2}{27n^2\rho^2 R} C \tag{143}$$

$$\implies \epsilon \leq \frac{l^4 c^2}{27n^2\rho^2 R} C \tag{144}$$

$$\iff \frac{27n^2\rho^2 R}{l^4 c^2}\epsilon \leq C, \tag{145}$$

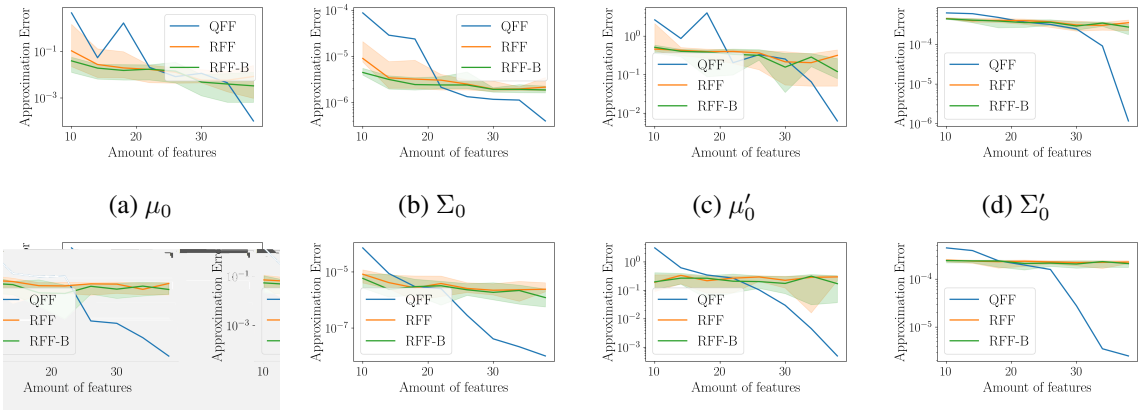
$$\tag{146}$$

which concludes the proof of this theorem.

The above bound could be reformulated as $m \geq 12 + \max(\frac{e}{2l^2}, \log(\frac{n^2\rho^3 R}{l^8 c^2 C}))$. Moreover, we assumed that $R \geq 1$ and that $c \leq 1$. If any of these conditions is not met, then the bounds are still valid if we substitute these quantities by 1 (the same holds also for ρ and l). Finally, we implicitly assumed that ϵ, the uniform upper bound of the approximation for $k_{\boldsymbol{\phi}}, {}' k_{\boldsymbol{\phi}}, k'_{\boldsymbol{\phi}}, k''_{\boldsymbol{\phi}}$ is smaller than 1. This happens when $m \geq 3 + \max(\frac{e}{2l^2}, \log(\frac{10\rho}{l^4}))$, a condition which is met if $m \geq 3 + \max(\frac{e}{2l^2}, \log(\frac{270n^2\rho^3 R}{l^8 c^2 C}))$.

Appendix D. Additional Empirical Evaluation GPR

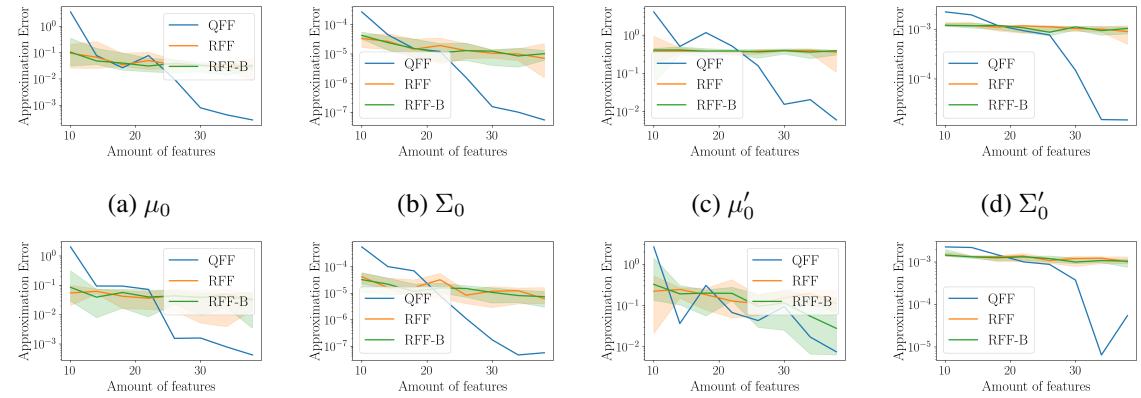
D.1 Lotka Volterra



(a) μ_0 (b) Σ_0 (c) μ_0' (d) Σ_0'

(e) μ_1 (f) Σ_1 (g) μ_1' (h) Σ_1'

Figure 8: Approximation error of the different feature approximations compared to the accurate GP, evaluated at $t = 1.75$ for the Lotka Volterra system with 1000 observations and $\sigma^2 = 0.1$. For each feature, we show the median as well as the 12.5% and 87.5% quantiles over 10 independent noise realizations, separately for each state dimension.

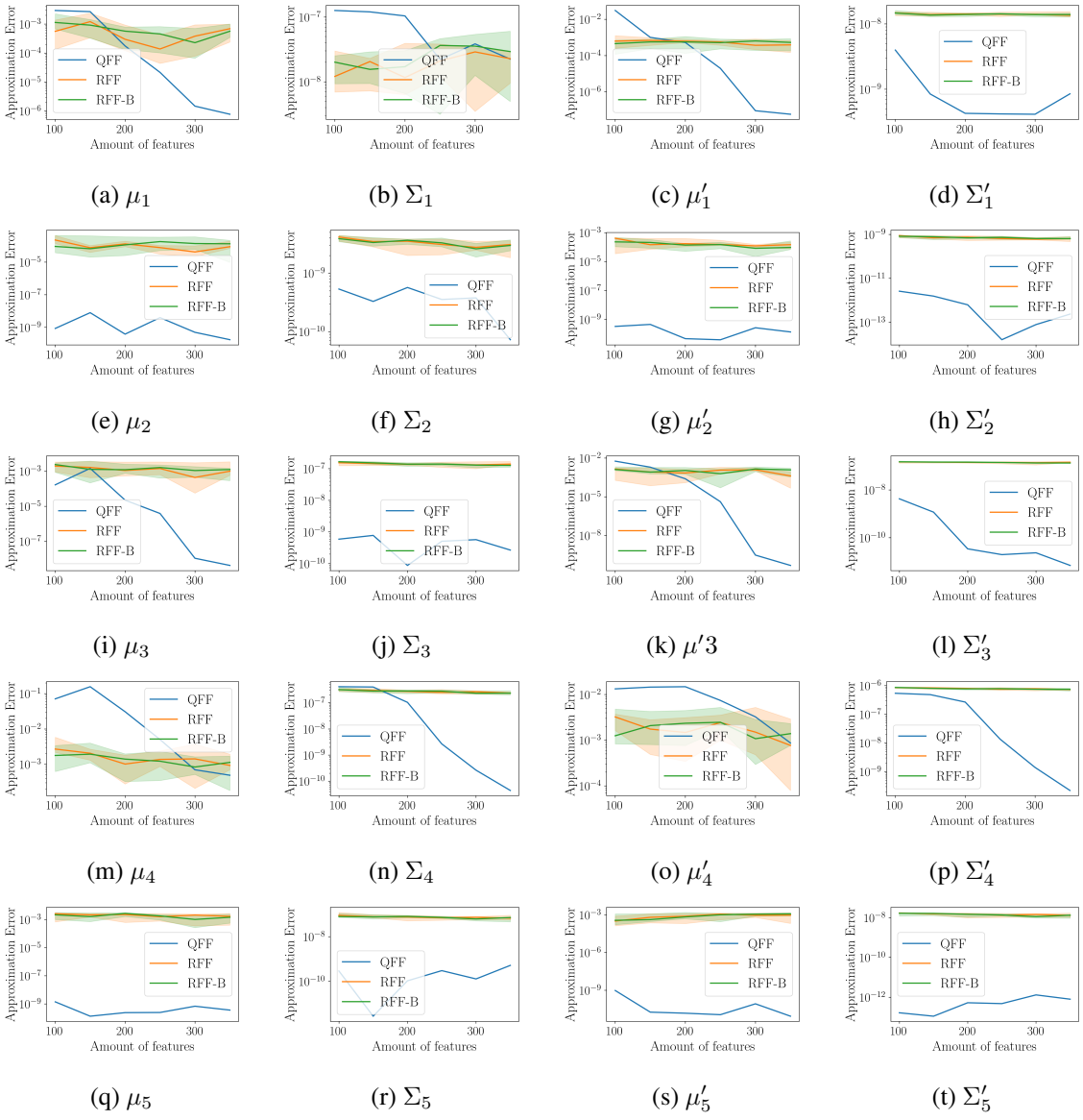


(a) μ_0 (b) Σ_0 (c) μ_0' (d) Σ_0'

(e) μ_1 (f) Σ_1 (g) μ_1' (h) Σ_1'

Figure 9: Approximation error of the different feature approximations compared to the accurate GP, evaluated at $t = 1.75$ for the Lotka Volterra system with 1000 observations and $\sigma^2 = 0.5$. For each feature, we show the median as well as the 12.5% and 87.5% quantiles over 10 independent noise realizations, separately for each state dimension.

D.2 Protein Transduction



(a) μ_1 (b) Σ_1 (c) μ_1' (d) Σ_1'

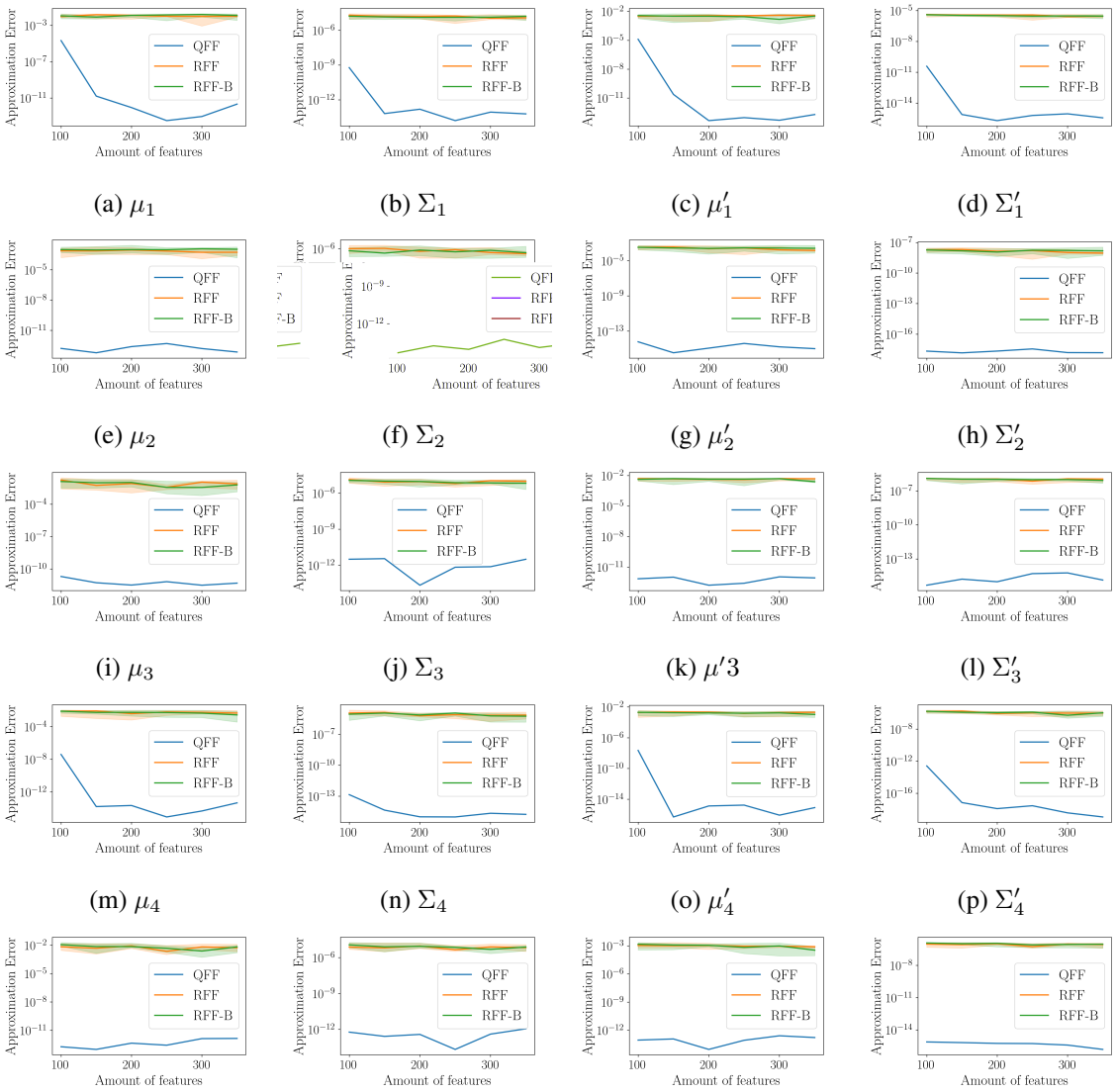
(e) μ_2 (f) Σ_2 (g) μ_2' (h) Σ_2'

(i) μ_3 (j) Σ_3 (k) $\mu'3$ (l) Σ_3'

(m) μ_4 (n) Σ_4 (o) μ_4' (p) Σ_4'

(q) μ_5 (r) Σ_5 (s) μ_5' (t) Σ_5'

Figure 10: Approximation error of the different feature approximations compared to the accurate GP, evaluated at $t = 30$ for the Protein Transduction system with 1000 observations and $\sigma^2 = 0.0001$. For each feature, we show the median as well as the 12.5% and 87.5% quantiles over 10 independent noise realizations, separately for each state dimension.



(a) μ_1 (b) Σ_1 (c) μ'_1 (d) Σ'_1

(e) μ_2 (f) Σ_2 (g) μ'_2 (h) Σ'_2

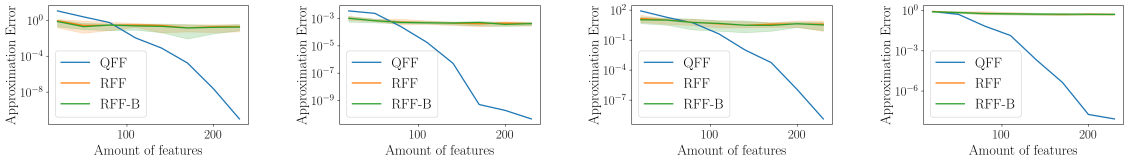
(i) μ_3 (j) Σ_3 (k) $\mu'3$ (l) Σ'_3

(m) μ_4 (n) Σ_4 (o) μ'_4 (p) Σ'_4

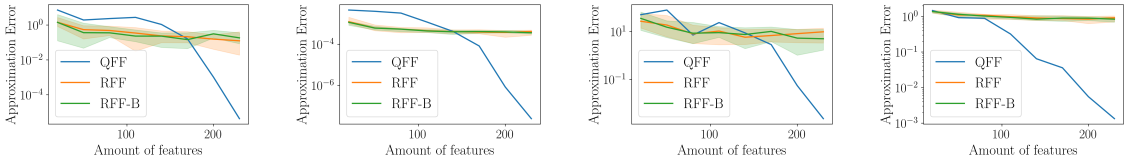
(q) μ_5 (r) Σ_5 (s) μ'_5 (t) Σ'_5

Figure 11: Approximation error of the different feature approximations compared to the accurate GP, evaluated at $t = 30$ for the Protein Transduction system with 1000 observations and $\sigma^2 = 0.01$. For each feature, we show the median as well as the 12.5% and 87.5% quantiles over 10 independent noise realizations, separately for each state dimension.

D.3 Lorenz

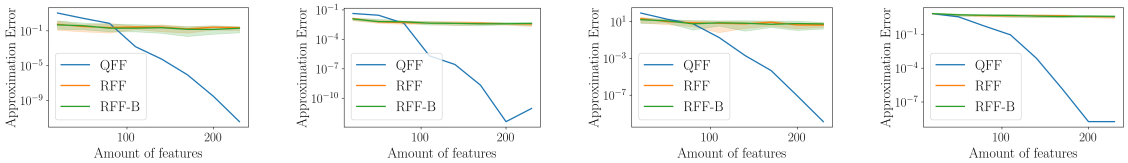


(a) μ_0 (b) Σ_0 (c) μ_0' (d) Σ_0'

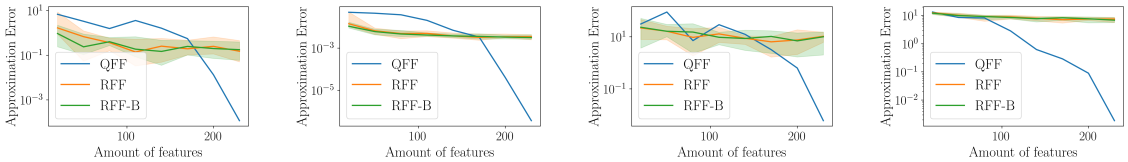


(e) μ_1 (f) Σ_1 (g) μ_1' (h) Σ_1'

Figure 12: Approximation error of the different feature approximations compared to the accurate GP, evaluated at $t = 0.8$ for the Lorenz system with 1000 observations and an SNR of 100. For each feature, we show the median as well as the 12.5% and 87.5% quantiles over 10 independent noise realizations, separately for each state dimension.

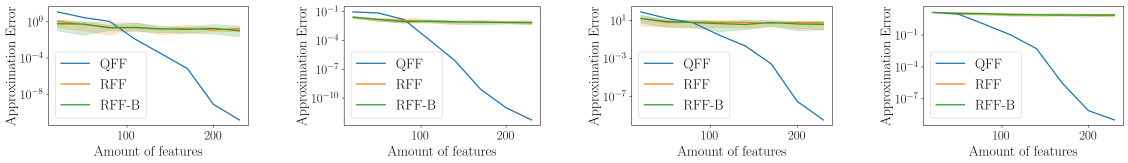


(a) μ_0 (b) Σ_0 (c) μ_0' (d) Σ_0'

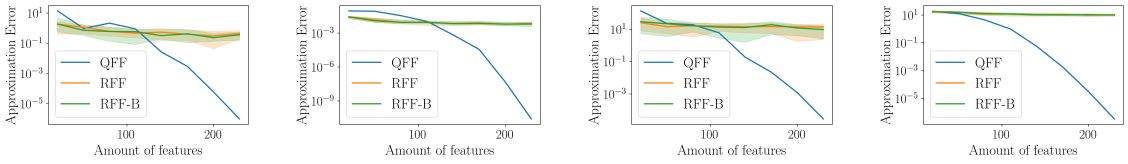


(e) μ_1 (f) Σ_1 (g) μ_1' (h) Σ_1'

Figure 13: Approximation error of the different feature approximations compared to the accurate GP, evaluated at $t = 0.8$ for the Lorenz system with 1000 observations and an SNR of 10. For each feature, we show the median as well as the 12.5% and 87.5% quantiles over 10 independent noise realizations, separately for each state dimension.



(a) μ_0 (b) Σ_0 (c) μ_0' (d) Σ_0'



(e) μ_1 (f) Σ_1 (g) μ_1' (h) Σ_1'

Figure 14: Approximation error of the different feature approximations compared to the accurate GP, evaluated at $t = 0.8$ for the Lorenz system with 1000 observations and an SNR of 5. For each feature, we show the median as well as the 12.5% and 87.5% quantiles over 10 independent noise realizations, separately for each state dimension.

Appendix E. Risk Approximation Error Bounds

Let

$$\mathcal{R}_{\lambda\gamma\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}) = \boldsymbol{x}^T(\boldsymbol{C}_{\boldsymbol{\phi}} + \lambda \boldsymbol{I})^{-1}\boldsymbol{x} \tag{147}$$

$$+ (\boldsymbol{x} - \boldsymbol{y})^T\sigma^{-2}(\boldsymbol{x} - \boldsymbol{y}) \tag{148}$$

$$+ (\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{x})^T(\boldsymbol{A} + \gamma \boldsymbol{I})^{-1}(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{x}) \tag{149}$$

where $\boldsymbol{C}_{\boldsymbol{\phi}}[i, j] = \rho e^{-\frac{r_{ij}^2}{2l^2}}$ for some fixed hyperparameters $\boldsymbol{\phi} = (\rho, l)$, which denote the variance and the lengthscale. Let n be the number of data points and m be the order of the Quadrature scheme used to approximate the kernel. By writing $\tilde{\boldsymbol{C}}_{\boldsymbol{\phi}}$, $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{D}}$ for the approximated quantities as described in Section **??**, we get get the approximate risk function

$$\tilde{\mathcal{R}}_{\lambda\gamma\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{\theta}) := \boldsymbol{x}^T(\tilde{\boldsymbol{C}}_{\boldsymbol{\phi}} + \lambda \boldsymbol{I})^{-1}\boldsymbol{x} \tag{150}$$

$$+ (\boldsymbol{x} - \boldsymbol{y})^T\sigma^{-2}(\boldsymbol{x} - \boldsymbol{y}) \tag{151}$$

$$+ (f(\boldsymbol{x}, \boldsymbol{\theta}) - \tilde{\boldsymbol{D}}\boldsymbol{x})^T(\tilde{\boldsymbol{A}} + \gamma \boldsymbol{I})^{-1}(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}) - \tilde{\boldsymbol{D}}\boldsymbol{x}). \tag{152}$$

To prove Theorem 4, we show that $\tilde{\mathcal{R}}$ converges to $\mathcal{R}$ in the relative error sense: As m increases, the relative error $\frac{|\mathcal{R}-\tilde{\mathcal{R}}|}{\mathcal{R}}$ will become arbitrarily small. Theorem 4 is restated here as Theorem 17.

Theorem 17 *Let $\mathcal{R}$ and $\tilde{\mathcal{R}}$ be defined as above. The parameters λ and γ, the kernel hyperparametes $\boldsymbol{\phi} = (\rho, l)$ and the number of data points n are considered fixed. We assume $n \geq 60$. Consider $1 > \epsilon > 0$. If m, the order of the quadrature scheme, is at least*

$$m \geq 10 + \max\{\frac{e}{2l^2}, \log_2(\frac{\rho^2 n^3}{\lambda^2 \gamma l^4 \epsilon})\} \tag{153}$$

then we have

$$\frac{|\mathcal{R}_{\lambda\gamma\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{\theta}) - \tilde{\mathcal{R}}_{\lambda\gamma\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{\theta})|}{\mathcal{R}_{\lambda\gamma\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{\theta})} \leq \epsilon \tag{154}$$

for any configuration of the variables $\boldsymbol{x}$ and $\boldsymbol{\theta}$.

The lengthscale corresponds to time observations normalized in the $[0,1]$ interval. In order to make things a bit less complicated in the calculations, we shall assume that $\gamma \leq 1$,$\lambda \leq 1$,$\rho \geq 1$, $l \leq 1$ (and specifically $l \leq \frac{e}{4}$). If any of these assumptions is violated, we can just substitute the corresponding parameter with 1 in the previous bound, and the resulting bound will still be valid. Moreover, the logarithm is the binary one and we will simply use $\log$ from now on. Before we prove the theorem, we will introduce some notations and some preliminary results.

Let $|\boldsymbol{K}|_F = \sqrt{\text{tr}(\boldsymbol{K}\boldsymbol{K}^T)}$ be the Frobenius norm of a matrix $\boldsymbol{K}$, $|\boldsymbol{K}|_{\max} = \max_{i,j}|\boldsymbol{K}_{ij}|$ the max norm of $\boldsymbol{K}$ and $\sigma_1(\boldsymbol{K}) := |\boldsymbol{K}|_2$ the spectral norm of $\boldsymbol{K}$ (which is by definition the largest singular value). It holds that $|\boldsymbol{K}\boldsymbol{x}| \leq \sigma_1(\boldsymbol{K})|\boldsymbol{x}|$. Moreover, for a $n \times n$ matrix $\boldsymbol{K}$ we have

$$\sigma_1(\boldsymbol{K}) \leq \sqrt{\sum \sigma_i^2(\boldsymbol{K})} = \sqrt{|\boldsymbol{K}|_F^2} \leq \sqrt{n^2|\boldsymbol{K}|_{max}^2} = n|\boldsymbol{K}|_{max} \tag{155}$$

Specifically, if $\boldsymbol{C}_{\boldsymbol{\phi}}[i,j] = \rho e^{-\frac{r_{ij}^2}{2l^2}}$, then

$$\sigma_1(\boldsymbol{C}_{\boldsymbol{\phi}}) \leq \rho n \tag{156}$$

while

$$\sigma_1(\boldsymbol{C}'_{\boldsymbol{\phi}}) \leq \frac{\rho}{l} n \tag{157}$$

since $\boldsymbol{C}'_{\boldsymbol{\phi}}[i,j] = -\frac{\rho}{l}\frac{r_{ij}}{l}e^{-\frac{r_{ij}^2}{2l^2}}$ as $xe^{-\frac{x^2}{2}} \leq \frac{1}{\sqrt{e}} \leq 1$.
From Woodbury's identity for matrix inversion, for invertible $\boldsymbol{K}$ and $\boldsymbol{K}+\boldsymbol{E}$ we have:

$$(\boldsymbol{K}+\boldsymbol{E})^{-1} = \boldsymbol{K}^{-1} - \boldsymbol{K}^{-1}(\boldsymbol{I}+\boldsymbol{E}\boldsymbol{K}^{-1})^{-1}\boldsymbol{E}\boldsymbol{K}^{-1} \tag{158}$$

Let $\boldsymbol{E}_1 := \boldsymbol{E}_1(m) = \tilde{\boldsymbol{C}}_{\boldsymbol{\phi}} - \boldsymbol{C}_{\boldsymbol{\phi}}$. We know that $|\boldsymbol{E}_1|_{max} \leq \sqrt{\frac{\pi}{2}}\frac{1}{m^m}(\frac{e}{4l^2})^m$ so for $a := \frac{e}{4l^2}$ we have

$$\sigma_1(\boldsymbol{E}_1) \leq 2(\frac{a}{m})^m n \tag{159}$$

Now let us bound the first term of the risk

Lemma 18 *Consider the term $\boldsymbol{x}^T(\boldsymbol{C}_{\boldsymbol{\phi}}+\lambda\boldsymbol{I})^{-1}\boldsymbol{x}$ approximated by $x^T(\tilde{\boldsymbol{C}}_{\boldsymbol{\phi}}+\lambda\boldsymbol{I})^{-1}x$. Then for $m \geq M := 7 + \max\{\frac{e}{2l^2}, \log(\frac{\rho^2 n^3}{\lambda^2\gamma l^4})\}$ we get*

$$|\boldsymbol{x}^T(\boldsymbol{C}_{\boldsymbol{\phi}}+\lambda\boldsymbol{I})^{-1}\boldsymbol{x} - \boldsymbol{x}^T(\tilde{\boldsymbol{C}}_{\boldsymbol{\phi}}+\lambda\boldsymbol{I})^{-1}\boldsymbol{x}| \leq \epsilon_1 \boldsymbol{x}^T(\boldsymbol{C}_{\boldsymbol{\phi}}+\lambda\boldsymbol{I})^{-1}\boldsymbol{x} \tag{160}$$

where $\epsilon_1 = \frac{4n}{\lambda}(\frac{a}{m})^m$.

Proof 7 *From Woodbury's inversion furmula (158), if we set $\boldsymbol{K} = \boldsymbol{C}_{\boldsymbol{\phi}} + \lambda\boldsymbol{I}$ and $\boldsymbol{E} = \boldsymbol{E}_1$ we have $\sigma_1(\boldsymbol{K}^{-1}) < \frac{1}{\lambda}$. Regarding the matrix $\boldsymbol{I}+\boldsymbol{E}\boldsymbol{K}^{-1}$, we would like to find a lower bound on the smallest singular value $\sigma_n(\boldsymbol{I}+\boldsymbol{E}\boldsymbol{K}^{-1})$. For any vector u with $|u|=1$ we have*

$$|(\boldsymbol{I}+\boldsymbol{E}\boldsymbol{K}^{-1})u| \geq |u| - |\boldsymbol{E}\boldsymbol{K}^{-1}u| \geq 1 - \sigma_1(\boldsymbol{E})\sigma_1(\boldsymbol{K}^{-1}) \geq 1 - \frac{2n}{\lambda}(\frac{a}{m})^m \geq \frac{1}{2} \tag{161}$$

for $m \geq \max\{\frac{e}{2l^2}, \log(\frac{4n}{\lambda})\}$, which is fullfiled if $m \geq M$. So $\sigma_n(\boldsymbol{I} + \boldsymbol{E}\boldsymbol{K}^{-1}) \geq \frac{1}{2}$ and consequently $\sigma_1((\boldsymbol{I}+\boldsymbol{E}\boldsymbol{K}^{-1})^{-1}) \leq 2$. Moreover, $\sigma_1(\boldsymbol{K}^{-1}(\boldsymbol{I}+\boldsymbol{E}\boldsymbol{K}^{-1})^{-1}\boldsymbol{E}\boldsymbol{K}^{-1}) \leq \sigma_1(\boldsymbol{K}^{-1})^2\sigma_1((\boldsymbol{I}+\boldsymbol{E}\boldsymbol{K}^{-1})^{-1})\sigma_1(\boldsymbol{E}) \leq \frac{4n}{\lambda^2}(\frac{a}{m})^m$. So

$$\sigma_1((\boldsymbol{C}_{\boldsymbol{\phi}} + \lambda\boldsymbol{I})^{-1} - (\tilde{\boldsymbol{C}}_{\boldsymbol{\phi}} + \lambda\boldsymbol{I})^{-1}) \leq \frac{4n}{\lambda^2}(\frac{a}{m})^m \tag{162}$$

In total we have

$$|\boldsymbol{x}^T(\boldsymbol{C}_{\boldsymbol{\phi}} + \lambda\boldsymbol{I})^{-1}\boldsymbol{x} - \boldsymbol{x}^T(\tilde{\boldsymbol{C}}_{\boldsymbol{\phi}} + \lambda\boldsymbol{I})^{-1}\boldsymbol{x}| = |\boldsymbol{x}^T\boldsymbol{K}^{-1}(\boldsymbol{I} + \boldsymbol{E}\boldsymbol{K}^{-1})^{-1}\boldsymbol{E}\boldsymbol{K}^{-1}\boldsymbol{x}| \leq \tag{163}$$

$$|\boldsymbol{x}^T\boldsymbol{K}^{-\frac{1}{2}}|\sigma_1(\boldsymbol{K}^{-\frac{1}{2}}(\boldsymbol{I} + \boldsymbol{E}\boldsymbol{K}^{-1})^{-1}\boldsymbol{E}\boldsymbol{K}^{-\frac{1}{2}})|\boldsymbol{x}^T\boldsymbol{K}^{-\frac{1}{2}}| \leq \tag{164}$$

$$|\boldsymbol{x}^T\boldsymbol{K}^{-\frac{1}{2}}||\boldsymbol{x}^T\boldsymbol{K}^{-\frac{1}{2}}|\sigma_1(\boldsymbol{K}^{-\frac{1}{2}})^2\sigma_1((\boldsymbol{I} + \boldsymbol{E}\boldsymbol{K}^{-1})^{-1})\sigma_1(\boldsymbol{E}) \leq \tag{165}$$

$$|\boldsymbol{x}^T(\boldsymbol{C}_{\boldsymbol{\phi}} + \lambda\boldsymbol{I})^{-1}\boldsymbol{x}|\frac{4n}{\lambda}(\frac{a}{m})^m \tag{166}$$

Now we define $\boldsymbol{E}_2 = \boldsymbol{E}_2(m) := \tilde{\boldsymbol{A}} - \boldsymbol{A}$ and $\boldsymbol{E}_3 = \boldsymbol{E}_3(m) := \tilde{\boldsymbol{D}} - \boldsymbol{D}$. We need to find an upper bound for the spectral norm of these two error terms.

Lemma 19 *Consider the error terms $\boldsymbol{E}_2$ and $\boldsymbol{E}_3$ as defined above. Then for $m \geq M := 7 + \max\{\frac{e}{2l^2}, \log(\frac{\rho^2n^3}{\lambda^2\gamma l^4})\}$ we get*

$$\sigma_1(\boldsymbol{E}_2(m+3)) \leq 20\frac{\rho^2n^3a^2}{\lambda^2}(\frac{a}{m})^m \quad \sigma_1(\boldsymbol{E}_3(m+3)) \leq 10\frac{n^2\rho a}{\lambda^2}(\frac{a}{m})^m \tag{167}$$

Proof 8 *We have*

$$\boldsymbol{E}_2 = \tilde{\boldsymbol{C}}_{\boldsymbol{\phi}}'' - \boldsymbol{C}_{\boldsymbol{\phi}}'' + {}'\tilde{\boldsymbol{C}}_{\boldsymbol{\phi}}(\tilde{\boldsymbol{C}}_{\boldsymbol{\phi}} + \lambda\boldsymbol{I})^{-1}\tilde{\boldsymbol{C}}_{\boldsymbol{\phi}}' - {}'\boldsymbol{C}_{\boldsymbol{\phi}}(\boldsymbol{C}_{\boldsymbol{\phi}} + \lambda\boldsymbol{I})^{-1}\boldsymbol{C}_{\boldsymbol{\phi}}' \tag{168}$$

and we define $\boldsymbol{E}_{21} := \tilde{\boldsymbol{C}}_{\boldsymbol{\phi}}' - \boldsymbol{C}_{\boldsymbol{\phi}}'$, $\boldsymbol{E}_{22} := (\tilde{\boldsymbol{C}}_{\boldsymbol{\phi}} + \lambda\boldsymbol{I})^{-1} - (\boldsymbol{C}_{\boldsymbol{\phi}} + \lambda\boldsymbol{I})^{-1}$ and $\boldsymbol{E}_{23} := \tilde{\boldsymbol{C}}_{\boldsymbol{\phi}}'' - \boldsymbol{C}_{\boldsymbol{\phi}}''$. We know that $|\boldsymbol{E}_{21}(m+2)|_{max} \leq 16a(\frac{a}{m})^m$ so $\sigma_1(\boldsymbol{E}_{21}(m+3)) \leq 16a(\frac{a}{m})^m n$. Moreover, we know $|\boldsymbol{E}_{23}(m+3)|_{max} \leq 32a^2(\frac{a}{m})^m$ so $\sigma_1(\boldsymbol{E}_{23}(m+3)) = 32a^2(\frac{a}{m})^m n$. Finally, $\sigma_1(E_{22}(m)) \leq \frac{4n}{\lambda^2}(\frac{a}{m})^m$ (by (162)). So

$$\sigma_1(\boldsymbol{E}_2(m+3)) = \tag{169}$$

$$\sigma_1(\boldsymbol{E}_{23} + ({}'\boldsymbol{C}_{\boldsymbol{\phi}} + \boldsymbol{E}_{21}^T)((\boldsymbol{C}_{\boldsymbol{\phi}} + \lambda\boldsymbol{I})^{-1} + \boldsymbol{E}_{22})(\boldsymbol{C}_{\boldsymbol{\phi}}' + \boldsymbol{E}_{21}) - {}'\boldsymbol{C}_{\boldsymbol{\phi}}(\boldsymbol{C}_{\boldsymbol{\phi}} + \lambda\boldsymbol{I})^{-1}\boldsymbol{C}_{\boldsymbol{\phi}}') \leq \tag{170}$$

$$\sigma_1(\boldsymbol{E}_{23}) + 2\sigma_1({}'\boldsymbol{C}_{\boldsymbol{\phi}})\sigma_1((\boldsymbol{C}_{\boldsymbol{\phi}} + \lambda\boldsymbol{I})^{-1})\sigma_1(\boldsymbol{E}_{21}) + \sigma_1(\boldsymbol{E}_{21})^2\sigma_1((\boldsymbol{C}_{\boldsymbol{\phi}} + \lambda\boldsymbol{I})^{-1}) + \tag{171}$$

$$\sigma_1({}'\boldsymbol{C}_{\boldsymbol{\phi}})^2\sigma_1(\boldsymbol{E}_{22}) + 2\sigma_1(\boldsymbol{E}_{21})\sigma_1(\boldsymbol{E}_{22})\sigma_1({}'\boldsymbol{C}_{\boldsymbol{\phi}}) + \sigma_1(\boldsymbol{E}_{21})^2\sigma_1(\boldsymbol{E}_{22}) \leq \tag{172}$$

$$20\frac{\rho^2n^3a^2}{\lambda^2}(\frac{a}{m})^m \tag{173}$$

Now we will prove for each summand in the last sum that it is at most $\frac{\rho^2 n^3 a^2}{\lambda^2}(\frac{a}{m})^m$ *times a constant. We will use the fact that* $m \geq M$*. Indeed:*

$$\sigma_1(\boldsymbol{E}_{23}) \leq 32a^2(\frac{a}{m})^m n \leq \frac{1}{2}\frac{\rho^2 n^3 a^2}{\lambda^2}(\frac{a}{m})^m \tag{174}$$

$$2\sigma_1('\boldsymbol{C}_{\boldsymbol{\phi}})\sigma_1((\boldsymbol{C}_{\boldsymbol{\phi}}+\lambda\boldsymbol{I})^{-1})\sigma_1(\boldsymbol{E}_{21}) \leq 32\frac{\rho a}{l\lambda}(\frac{a}{m})^m n^2 \leq \frac{\rho^2 n^3 a^2}{\lambda^2}(\frac{a}{m})^m \tag{175}$$

$$\sigma_1(\boldsymbol{E}_{21})^2\sigma_1((\boldsymbol{C}_{\boldsymbol{\phi}}+\lambda\boldsymbol{I})^{-1}) \leq 16^2\frac{a^2n^2}{\lambda}(\frac{a}{m})^{2m} \leq \frac{1}{2}\frac{\rho^2 n^3 a^2}{\lambda^2}(\frac{a}{m})^m \tag{176}$$

$$\sigma_1('\boldsymbol{C}_{\boldsymbol{\phi}})^2\sigma_1(\boldsymbol{E}_{22}) \leq 16\frac{\rho^2 n^3 a}{\lambda^2}(\frac{a}{m})^m \leq 16\frac{\rho^2 n^3 a^2}{\lambda^2}(\frac{a}{m})^m \tag{177}$$

$$2\sigma_1(\boldsymbol{E}_{21})\sigma_1(\boldsymbol{E}_{22})\sigma_1('\boldsymbol{C}_{\boldsymbol{\phi}}) \leq 16a(\frac{a}{m})^m n\frac{4n}{\lambda^2}(\frac{a}{m})^m\frac{n\rho}{l} \leq 2^9\frac{a^2n^2\rho^2}{\lambda^2}(\frac{a}{m})^{2m} \leq \frac{\rho^2 n^3 a^2}{\lambda^2}(\frac{a}{m})^m \tag{178}$$

$$\sigma_1(\boldsymbol{E}_{21})^2\sigma_1(\boldsymbol{E}_{22}) \leq 2^{10}\frac{a^2n^3}{\lambda^2}(\frac{a}{m})^{3m} \leq \frac{\rho^2 n^3 a^2}{\lambda^2}(\frac{a}{m})^m \tag{179}$$

Similarly,

$$\boldsymbol{E}_3 = '\tilde{\boldsymbol{C}}_{\boldsymbol{\phi}}(\tilde{\boldsymbol{C}}_{\boldsymbol{\phi}}+\lambda\boldsymbol{I})^{-1} - '\boldsymbol{C}_{\boldsymbol{\phi}}(\boldsymbol{C}_{\boldsymbol{\phi}}+\lambda\boldsymbol{I})^{-1} = '\boldsymbol{C}_{\boldsymbol{\phi}}\boldsymbol{E}_{22} + \boldsymbol{E}_{21}^T(\boldsymbol{C}_{\boldsymbol{\phi}}+\lambda\boldsymbol{I})^{-1} + \boldsymbol{E}_{21}^T\boldsymbol{E}_{22} \tag{180}$$

Using again that $m \geq M$ *we have:*

$$\sigma_1('\boldsymbol{C}_{\boldsymbol{\phi}}\boldsymbol{E}_{22}) \leq \frac{4n^2\rho}{\lambda^2 l}(\frac{a}{m})^m \leq 8\frac{n^2\rho a}{\lambda^2}(\frac{a}{m})^m \tag{181}$$

$$\sigma_1(\boldsymbol{E}_{21}^T(\boldsymbol{C}_{\boldsymbol{\phi}}+\lambda\boldsymbol{I})^{-1}) \leq 16\frac{n}{\lambda}a(\frac{a}{m})^m \leq \frac{n^2\rho a}{\lambda^2}(\frac{a}{m})^m \tag{182}$$

$$\sigma(\boldsymbol{E}_{21}^T\boldsymbol{E}_{22}) \leq 16na(\frac{a}{m})^m\frac{4n}{\lambda^2}(\frac{a}{m})^m = 64\frac{n^2 a}{\lambda^2}(\frac{a}{m})^{2m} \leq \frac{n^2\rho a}{\lambda^2}(\frac{a}{m})^m \tag{183}$$

$$\tag{184}$$

Thus, we have shown that $\sigma_1(\boldsymbol{E}_3) \leq 10\frac{n^2\rho a}{\lambda^2}(\frac{a}{m})^m$

For the third term, $(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})^T(\boldsymbol{A}+\gamma\boldsymbol{I})^{-1}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})$ we have

$$|(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\tilde{\boldsymbol{D}}\boldsymbol{x})^T(\tilde{\boldsymbol{A}}+\gamma\boldsymbol{I})^{-1}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\tilde{\boldsymbol{D}}\boldsymbol{x}) - (\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})^T(\boldsymbol{A}+\gamma\boldsymbol{I})^{-1}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})| = \tag{185}$$

$$|(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\tilde{\boldsymbol{D}}\boldsymbol{x})^T(\tilde{\boldsymbol{A}}+\gamma\boldsymbol{I})^{-1}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\tilde{\boldsymbol{D}}\boldsymbol{x}) - (\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})^T(\tilde{\boldsymbol{A}}+\gamma\boldsymbol{I})^{-1}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})+ \tag{186}$$

$$(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})^T(\tilde{\boldsymbol{A}}+\gamma\boldsymbol{I})^{-1}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x}) - (\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})^T(\boldsymbol{A}+\gamma\boldsymbol{I})^{-1}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})| \leq \tag{187}$$

$$|(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\tilde{\boldsymbol{D}}\boldsymbol{x})^T(\tilde{\boldsymbol{A}}+\gamma\boldsymbol{I})^{-1}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\tilde{\boldsymbol{D}}\boldsymbol{x}) - (\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})^T(\tilde{\boldsymbol{A}}+\gamma\boldsymbol{I})^{-1}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})|+ \tag{188}$$

$$|(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})^T(\tilde{\boldsymbol{A}}+\gamma\boldsymbol{I})^{-1}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x}) - (\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})^T(\boldsymbol{A}+\gamma\boldsymbol{I})^{-1}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})| \tag{189}$$

$$:= T_1 + T_2 \tag{190}$$

This sum can be bounded by bounding each summand separately.

We start with T_2:

Lemma 20 *Consider the term T_2 as defined above. Then for $m \geq M := 7 + \max\{\frac{e}{2l^2}, \log(\frac{\rho^2 n^3}{\lambda^2 \gamma l^4})\}$, if we use for the approximations quadrature schemes of order $m+3$ we have:*

$$T_2 \leq \epsilon_{32} (\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{x})^T (\boldsymbol{A} + \gamma \boldsymbol{I})^{-1} (\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{x}) \tag{191}$$

where

$$\epsilon_{32} = 40 \frac{\rho^2 n^3 a^2}{\lambda^2 \gamma} (\frac{a}{m})^m \tag{192}$$

Proof 9 *As in the proof of Lemma 18, we can use Woodbury's identity. Now, we use $\boldsymbol{A} + \gamma \boldsymbol{I}$ instead of $\boldsymbol{C_\phi} + \lambda \boldsymbol{I}$, $\boldsymbol{E}_2$ instead of $\boldsymbol{E}_1$ and $(\boldsymbol{f}(x, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{x})$ instead of $\boldsymbol{x}$. This directly yields*

$$\sigma_1(\boldsymbol{K}^{-\frac{1}{2}})^2 \sigma_1((\boldsymbol{I} + \boldsymbol{E}\boldsymbol{K}^{-1})^{-1}) \sigma_1(\boldsymbol{E}_2) \leq 40 \frac{\rho^2 n^3 a^2}{\lambda^2 \gamma} (\frac{a}{m})^m. \tag{193}$$

Lemma 21 *Consider the term T_1 as defined above. Then for $m \geq M := 7 + \max\{\frac{e}{2l^2}, \log(\frac{\rho^2 n^3}{\lambda^2 \gamma l^4})\}$, if we use for the approximations quadrature schemes of order $m+3$ we have:*

$$T_1 \leq \epsilon_{31} (\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{x})^T (\boldsymbol{A} + \gamma \boldsymbol{I})^{-1} (\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{x}) + \epsilon_{31} \boldsymbol{x}^T (\boldsymbol{C_\phi} + \lambda \boldsymbol{I})^{-1} \boldsymbol{x} \tag{194}$$

where $\epsilon_{31} = 30 \frac{n^{\frac{5}{2}} \rho^{\frac{3}{2}} a}{\lambda^2 \gamma^{\frac{1}{2}}} (\frac{a}{m})^m$.

Proof 10 *It holds that*

$$\sigma_1((\tilde{\boldsymbol{A}} + \gamma \boldsymbol{I})^{-\frac{1}{2}}) \sigma_1(\boldsymbol{E}_3) \sigma_1((\boldsymbol{C_\phi} + \lambda \boldsymbol{I})^{\frac{1}{2}}) \leq 10 \frac{n^2 \rho a}{\lambda^2} (\frac{a}{m})^m (\frac{\rho n + \lambda}{\gamma})^{\frac{1}{2}} \leq \tag{195}$$

$$15 \frac{n^{\frac{5}{2}} \rho^{\frac{3}{2}} a}{\lambda^2 \gamma^{\frac{1}{2}}} (\frac{a}{m})^m = \frac{\epsilon_{31}}{2} \tag{196}$$

Simply using that $\boldsymbol{E}_3 = \tilde{\boldsymbol{D}} - \boldsymbol{D}$ we get:

$$T_1 = \tag{197}$$

$$|(\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}) - \tilde{\boldsymbol{D}}\boldsymbol{x})^T (\tilde{\boldsymbol{A}} + \gamma \boldsymbol{I})^{-1} (\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}) - \tilde{\boldsymbol{D}}\boldsymbol{x}) - (\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{x})^T (\tilde{\boldsymbol{A}} + \gamma \boldsymbol{I})^{-1} (\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{x})| = \tag{198}$$

$$|2(\boldsymbol{E}_3 \boldsymbol{x})^T (\tilde{\boldsymbol{A}} + \gamma \boldsymbol{I})^{-1} (\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{x}) + (\boldsymbol{E}_3 \boldsymbol{x})^T (\tilde{\boldsymbol{A}} + \gamma \boldsymbol{I})^{-1} (\boldsymbol{E}_3 \boldsymbol{x})| \leq \tag{199}$$

$$2|\boldsymbol{E}_3 \boldsymbol{x}| \sigma_1((\tilde{\boldsymbol{A}} + \gamma \boldsymbol{I})^{-\frac{1}{2}}) |(\tilde{\boldsymbol{A}} + \gamma \boldsymbol{I})^{-\frac{1}{2}} (\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}) - \boldsymbol{D}\boldsymbol{x})| + \sigma_1((\tilde{\boldsymbol{A}} + \gamma \boldsymbol{I})^{-1}) |\boldsymbol{E}_3 \boldsymbol{x}|^2 \tag{200}$$

Since $|\boldsymbol{E}_3 \boldsymbol{x}| = |\boldsymbol{E}_3 (\boldsymbol{C_\phi} + \lambda \boldsymbol{I})^{\frac{1}{2}} (\boldsymbol{C_\phi} + \lambda \boldsymbol{I})^{-\frac{1}{2}} \boldsymbol{x}| \leq \sigma_1(\boldsymbol{E}_3) \sigma_1((\boldsymbol{C_\phi} + \lambda \boldsymbol{I})^{\frac{1}{2}}) |(\boldsymbol{C_\phi} + \lambda \boldsymbol{I})^{-\frac{1}{2}} \boldsymbol{x}|$ we have:

$$\sigma_1((\tilde{\boldsymbol{A}} + \gamma \boldsymbol{I})^{-1}) |\boldsymbol{E}_3 \boldsymbol{x}|^2 \leq \sigma_1((\tilde{\boldsymbol{A}} + \gamma \boldsymbol{I})^{-1}) \sigma_1(\boldsymbol{E}_3)^2 \sigma_1((\boldsymbol{C_\phi} + \lambda \boldsymbol{I})^{\frac{1}{2}})^2 |(\boldsymbol{C_\phi} + \lambda \boldsymbol{I})^{-\frac{1}{2}} \boldsymbol{x}|^2 \leq \tag{201}$$

$$\frac{\epsilon_{31}^2}{4} |(\boldsymbol{C_\phi} + \lambda \boldsymbol{I})^{-\frac{1}{2}} \boldsymbol{x}|^2 \leq \frac{\epsilon_{31}}{2} |(\boldsymbol{C_\phi} + \lambda \boldsymbol{I})^{-\frac{1}{2}} \boldsymbol{x}|^2 \tag{202}$$

Note that $|(\boldsymbol{C_\phi}+\lambda\boldsymbol{I})^{-\frac{1}{2}}\boldsymbol{x}|^2$ is the first term of $\mathcal{R}_{\lambda\gamma\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{\theta})$
Moreover

$$2|\boldsymbol{E}_3\boldsymbol{x}|\sigma_1((\tilde{\boldsymbol{A}}+\gamma\boldsymbol{I})^{-\frac{1}{2}})|(\tilde{\boldsymbol{A}}+\gamma\boldsymbol{I})^{-\frac{1}{2}}(f(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})| = \tag{203}$$

$$\sigma_1((\tilde{\boldsymbol{A}}+\gamma\boldsymbol{I})^{-\frac{1}{2}})2|\boldsymbol{E}_3(\boldsymbol{C_\phi}+\lambda\boldsymbol{I})^{\frac{1}{2}}(\boldsymbol{C_\phi}+\lambda\boldsymbol{I})^{-\frac{1}{2}}\boldsymbol{x}||(\tilde{\boldsymbol{A}}+\gamma\boldsymbol{I})^{-\frac{1}{2}}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})| \leq \tag{204}$$

$$\sigma_1((\tilde{\boldsymbol{A}}+\gamma\boldsymbol{I})^{-\frac{1}{2}})\sigma_1(\boldsymbol{E}_3)\sigma_1((\boldsymbol{C_\phi}+\lambda\boldsymbol{I})^{\frac{1}{2}})2|(\boldsymbol{C_\phi}+\lambda\boldsymbol{I})^{-\frac{1}{2}}\boldsymbol{x}||(\tilde{\boldsymbol{A}}+\gamma\boldsymbol{I})^{-\frac{1}{2}}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})| \leq \tag{205}$$

$$\frac{\epsilon_{31}}{2}(|(\boldsymbol{C_\phi}+\lambda\boldsymbol{I})^{-\frac{1}{2}}\boldsymbol{x}|^2+|(\tilde{\boldsymbol{A}}+\gamma\boldsymbol{I})^{-\frac{1}{2}}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})|^2) \leq \tag{206}$$

$$\frac{\epsilon_{31}}{2}(|(\boldsymbol{C_\phi}+\lambda\boldsymbol{I})^{-\frac{1}{2}}\boldsymbol{x}|^2+(1+\epsilon_{32})|(\boldsymbol{A}+\gamma\boldsymbol{I})^{-\frac{1}{2}}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})|^2) \quad \textit{from (191)} \leq \tag{207}$$

$$\frac{\epsilon_{31}}{2}|(\boldsymbol{C_\phi}+\lambda\boldsymbol{I})^{-\frac{1}{2}}\boldsymbol{x}|^2+\epsilon_{31}|(\boldsymbol{A}+\gamma\boldsymbol{I})^{-\frac{1}{2}}(\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{D}\boldsymbol{x})|^2 \quad (\epsilon_{32}\leq 1 \textit{ for } m\geq M) \tag{208}$$

Combining all of these pieces, we can now proof the original Theorem.

Proof 11 (Proof of Theorem 17) *Consider $m \geq M := 7+\max\{\frac{e}{2l^2},\log(\frac{\rho^2n^3}{\lambda^2\gamma l^4})\}$ and let $m' = m+3$. Then, if we apply quadrature schemes of order m' for $\tilde{\mathcal{R}}$,using the results from Lemmata ((18), (21) and (20)) and that $\epsilon_1 \leq \epsilon_{32}$ and $\epsilon_{31} \leq \epsilon_{32}$ we get*

$$\frac{|\mathcal{R}_{\lambda\gamma\boldsymbol{\phi}}(x,\boldsymbol{\theta})-\tilde{\mathcal{R}}_{\lambda\gamma\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{\theta})|}{\mathcal{R}_{\lambda\gamma\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{\theta})} \leq 2\epsilon_{32} = 80\frac{\rho^2n^3a^2}{\lambda^2\gamma}(\frac{a}{m})^m \leq 50\frac{\rho^2n^3}{\lambda^2\gamma l^4}(\frac{a}{m})^m \tag{209}$$

In order to make that smaller than ϵ it suffices $m \geq \max\{\frac{e}{2l^2},\log(50\frac{\rho^2n^3}{\lambda^2\gamma l^4\epsilon})\}$. So we can choose for the order of the quadrature scheme m' to be

$$m' = 10+\max\{\frac{e}{2l^2},\log(\frac{\rho^2n^3}{\lambda^2\gamma l^4\epsilon})\} \tag{210}$$

Appendix F. Experimental Setups

Here, we will give a brief overview of the experimental setups we used to evaluate ODIN-S.

F.1 Basic Definitions

The trajectory RMSE has proven to be an efficient metric to evaluate the quality of a parameter inference scheme, especially in the context of non-identifiable systems. Here, we restate its definition exactly as provided by Wenk et al. (2020).

Definition 22 (Trajectory RMSE) *Let $\hat{\boldsymbol{\theta}}$ be the parameters estimated by an algorithm. Let $\mathbf{t}$ be the vector collecting the observation times. Define $\tilde{\mathbf{x}}(t)$ as the trajectory one obtains by integrating the ODEs using the estimated parameters, but the true initial value, i.e.*

$$\tilde{x}(0) = \mathbf{x}^*(0) \tag{211}$$

$$\tilde{x}(t) = \int_0^t f(\tilde{x}(s), \hat{\boldsymbol{\theta}}) ds \tag{212}$$

and define $\tilde{\mathbf{x}}$ element-wise as its evaluation at observation times $\mathbf{t}$, i.e. $\tilde{\mathbf{x}}_i = \tilde{x}(t_i)$. The trajectory RMSE is then defined as

$$\textit{tRMSE} \coloneqq \frac{1}{N} ||\tilde{\mathbf{x}} - \mathbf{x}||_2 \tag{213}$$

where $||.||_2$ denotes the standard Euclidean 2-norm.

Additionally, we would like to restate the definition of the signal-to-noise ratio, as it was used in our work to create the observation noise for the Quadrocopter system.

Definition 23 (Signal to Noise Ratio) *Let $x(t)$ be a time-continuous signal for a closed time interval. Let σ_x^2 denote its variance across time. Furthermore, let σ^2 be the variance of an additive, Gaussian noise signal. Then we define the SNR as the ration of these two variances, i.e.*

$$\textit{SNR} = \frac{\sigma_x^2}{\sigma^2}. \tag{214}$$

F.2 Lotka Volterra

$$\begin{aligned} \dot{x}_1(t) = &\quad \theta_1 x_1(t) - \theta_2 x_1(t) x_2(t) \\ \dot{x}_2(t) = &-\theta_3 x_2(t) + \theta_4 x_1(t) x_2(t). \end{aligned} \tag{215}$$

The Lotka Volterra system (Lotka, 1932) has become a widely used benchmarking system. Due to its locally linear dynamics (Gorbach et al., 2017) and relatively tame trajectories, it is a system many algorithms can solve. We follow the standard setting in the literature and use $\boldsymbol{\theta} = [2, 1, 4, 1]$ and $\mathbf{x}(0) = [5, 3]$ to generate trajectories over the time interval $[0, 2]$. The dynamics are shown in Equation (215).

F.3 Protein Transduction

$$\begin{aligned}
\dot{S} &= -\theta_1 S - \theta_2 SR + \theta_3 R_S \\
\dot{dS} &= \theta_1 S \\
\dot{R} &= -\theta_2 SR + \theta_3 R_S + \theta_5 \frac{R_{pp}}{\theta_6 + R_{pp}} \\
\dot{R}_S &= \theta_2 SR - \theta_3 R_S - \theta_4 R_S \\
\dot{R}_{pp} &= \theta_4 R_S - \theta_5 \frac{R_{pp}}{\theta_6 + R_{pp}}
\end{aligned} \tag{216}$$

A more challenging system was introduced by Vyshemirsky and Girolami (2007). Its nonlinear terms and non-stationarity introduce interesting challenges for many collocation methods. We follow the standard setting in the literature and use $\boldsymbol{\theta} = [0.07, 0.6, 0.05, 0.3, 0.017, 0.3]$ and $\mathbf{x}(0) = [1, 0, 1, 0, 0]$, but change the time interval to generate trajectories over the time interval $[0, 50]$, since they stay pretty much constant for $t > 50$. The dynamics are shown in Equation (216).

F.4 Lorenz 63

$$\dot{x} = \theta_0(y - x) \tag{217}$$
$$\dot{y} = x(\theta_1 - z) - y \tag{218}$$
$$\dot{z} = xy - \theta_2 z \tag{219}$$

The Lorenz 63 system was introduced by Lorenz (1963) to model atmospheric flows. It is an optimal test bed for parameter inference algorithms, as it exhibits chaotic behavior for the parameter settings we chose. Working with chaotic dynamics is notoriously challenging due to high sensitivity to parameter changes and the presence of many local optima. We follow standard literature and use $\boldsymbol{\theta} = [10, 28, 8/3]$ and $\mathbf{x}(0) = [1, 1, 1]$ to generate trajectories over the time interval $[0, 1]$. The dynamics are shown in Equation (219).

F.5 Quadrocopter

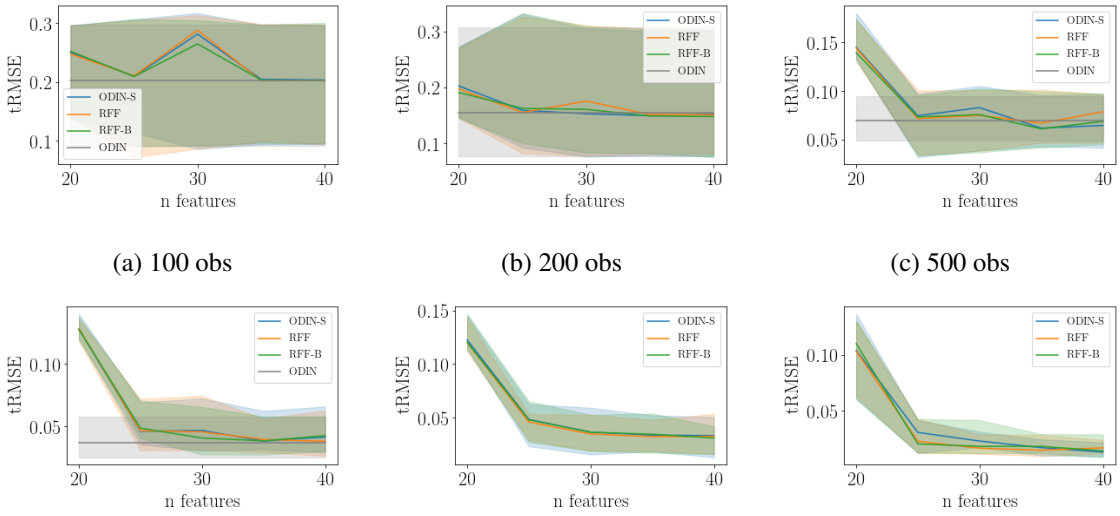
$$
\begin{aligned}
\dot{x}_0 &= -g\sin(x_7) + x_5x_1 - x_4x_2 \\
\dot{x}_1 &= g\sin(x_6)\cos(x_7) - x_0x_5 + x_2x_3 \\
\dot{x}_2 &= -\frac{u_0 + u_1 + u_2 + u_3}{\theta_0} + g\cos(x_6)\cos(x_7) + x_0x_4 - \theta_4x_1 \\
\dot{x}_3 &= \frac{1}{\theta_1}(\theta_5(-u_0 + u_1 + u_2 - u_3)) + (\theta_2 - \theta_3(\theta_2 + \theta_1)))x_4x_5) \\
\dot{x}_4 &= \frac{1}{\theta_2}(\theta_4(u_0 - u_1 + u_3 - u_4) + (\theta_3(\theta_2 + \theta_1) - \theta_1)x_3x_5) \\
\dot{x}_5 &= \frac{(\theta_1 - \theta_2)x_3x_4}{\theta_3(\theta_2 + \theta_1)} \\
\dot{x}_6 &= x_3 + (x_4\sin(x_6) + \frac{x_5\cos(x_6)\sin(x_7)}{\cos(x_7)} \\
\dot{x}_7 &= x_4\cos(x_6) - x_5\sin(x_6) \\
\dot{x}_8 &= \frac{x_4\sin(x_6) + x_5\cos(x_6)}{\cos(x_7)} \\
\dot{x}_9 &= \cos(x_7)\cos(x_8)x_0 + (-\cos(x_6)\sin(x_8) + \sin(x_6)\sin(x_7)\cos(x_8))x_1 \\
&\quad + (\sin(x_6)\sin(x_8) + \cos(x_6)\sin(x_7)\cos(x_8))x_2 \\
\dot{x}_{10} &= \cos(x_7)\sin(x_8)x_0 + (\cos(x_6)\cos(x_8) + \sin(x_6)\sin(x_7)\sin(x_8))x_1 \\
&\quad + (\cos(x_6)\sin(x_7)\sin(x_8) - \sin(x_6)\cos(x_8))x_2 \\
\dot{x}_{11} &= \sin(x_7)x_0 - \sin(x_6)\cos(x_7)x_1 - \cos(x_6)\cos(x_7)x_2
\end{aligned} \tag{220}
$$

As an ultimate benchmark, we introduce a parametric model describing the dynamics of a 6DOF quadrocopter, shown in Equation (220). Its strongly nonlinear dynamics and the presence of inputs make it a formidable challenge. The states of this system are representing the linear velocities (x_0, x_1, x_2), the angular velocities (x_3, x_4, x_5), the angles (x_6, x_7, x_8) and the position (x_9, x_{10}, x_{11}) of the quadrocopter. The four inputs represent the forces applied at the four different propellers. While in principle any input commands could be incorporated, we keep the inputs constant at $u = [0.248, 0.2475, 0.24775, 0.24775]$. This input leads to interesting nonstationary climbing, pitching and rolling behavior. We use $\boldsymbol{\theta} = [0.1, 0.00062, 0.00113, 0.9, 0.114, 0.0825, 9.85]$ and $x_i(0) = 0$ for $i = 0 \ldots 11$ to generate trajectories over the time interval $[0, 15]$.

Appendix G. Additional Empirical Evaluation SLEIPNIR

G.1 tRMSE vs Features

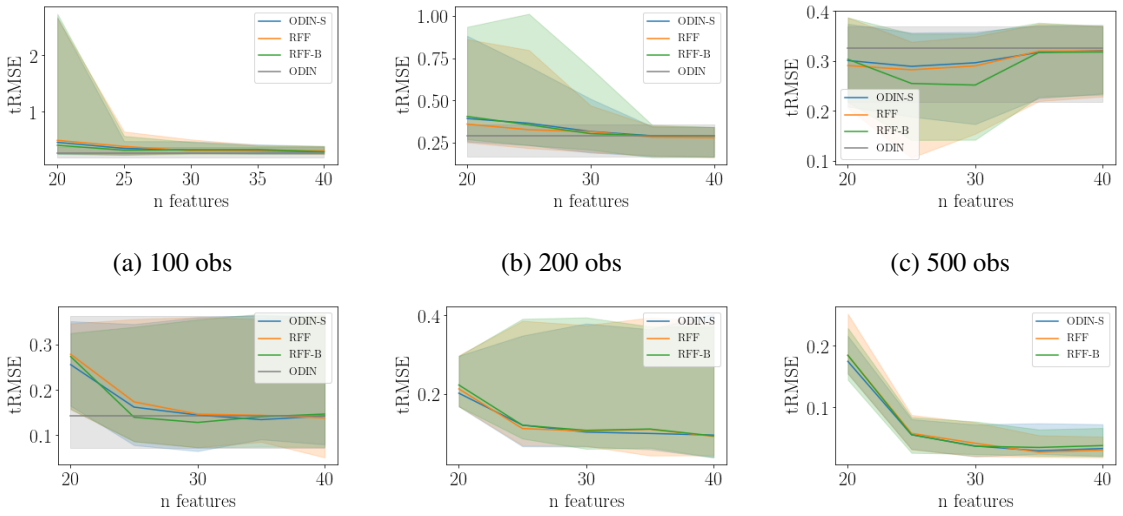
G.1.1 LOTKA VOLTERRA



(a) 100 obs (b) 200 obs (c) 500 obs

(d) 1000 obs (e) 2000 obs (f) 5000 obs

Figure 15: tRMSE vs features for the Lotka Volterra system using additive observation noise with $\sigma^2 = 0.1$.

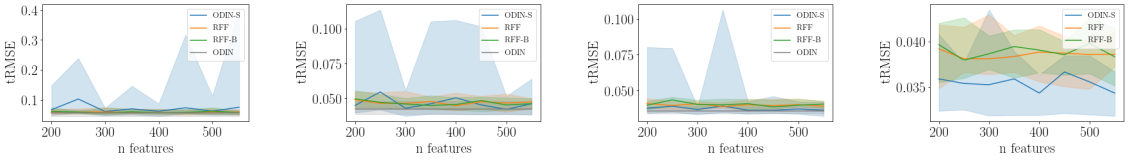


(a) 100 obs (b) 200 obs (c) 500 obs

(d) 1000 obs (e) 2000 obs (f) 5000 obs

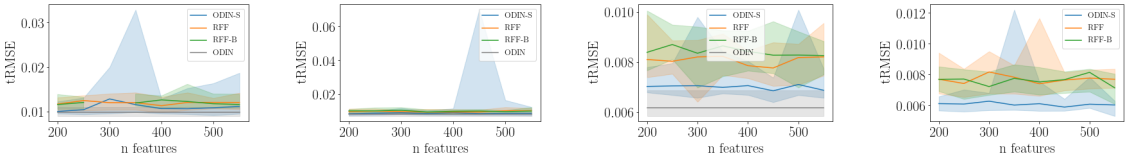
Figure 16: tRMSE vs features for the Lotka Volterra system using additive observation noise with $\sigma^2 = 0.5$.

G.1.2 PROTEIN TRANSDUCTION



(a) 500 obs (b) 1000 obs (c) 2000 obs (d) 3000 obs

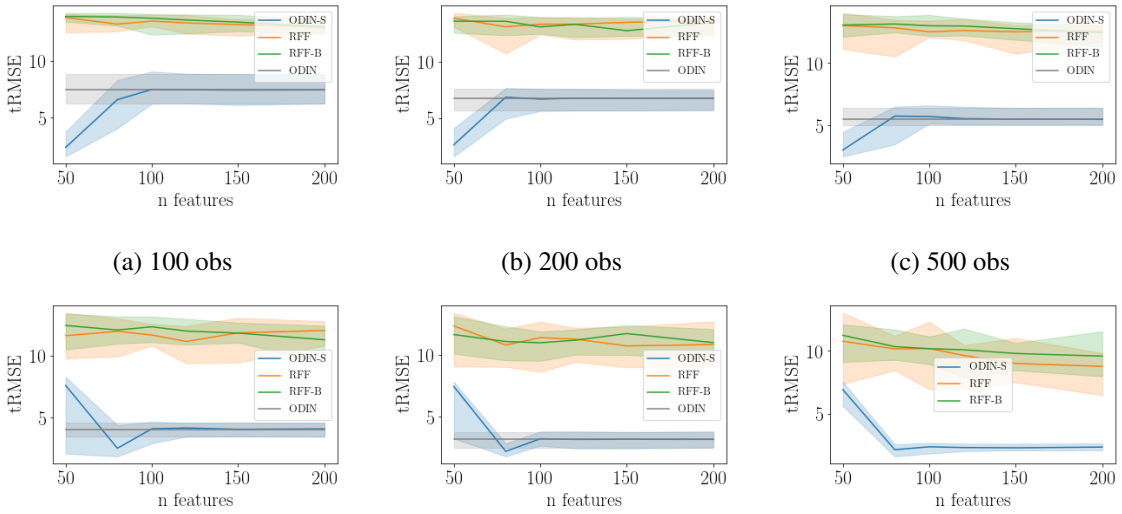
Figure 17: tRMSE vs features for the Protein Transduction system using additive observation noise with $\sigma^2 = 0.01$.



(a) 500 obs (b) 1000 obs (c) 2000 obs (d) 3000 obs

Figure 18: tRMSE vs features for the Protein Transduction system using additive observation noise with $\sigma^2 = 0.0001$.

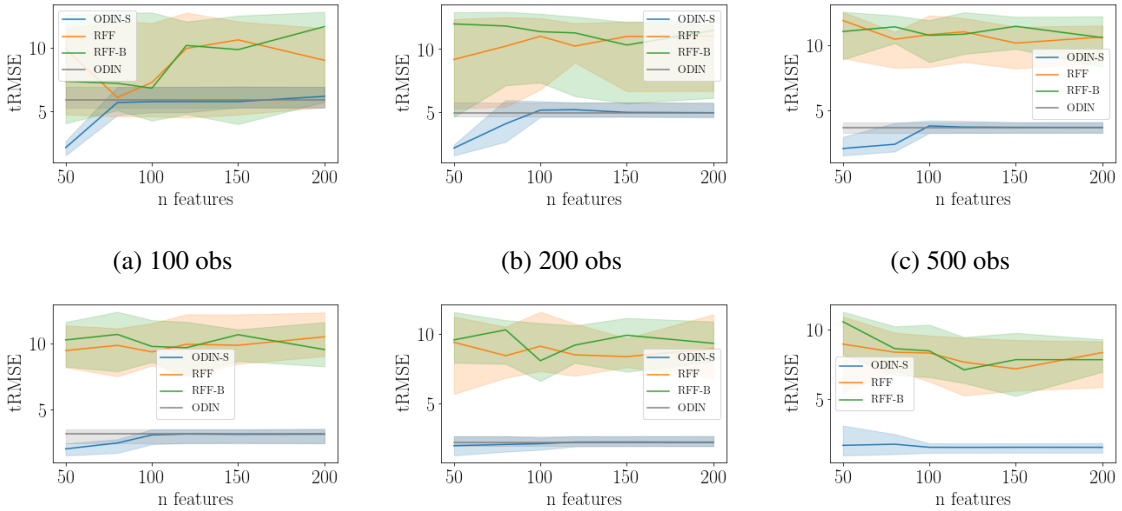
G.1.3 LORENZ



(a) 100 obs (b) 200 obs (c) 500 obs

(d) 1000 obs (e) 2000 obs (f) 5000 obs

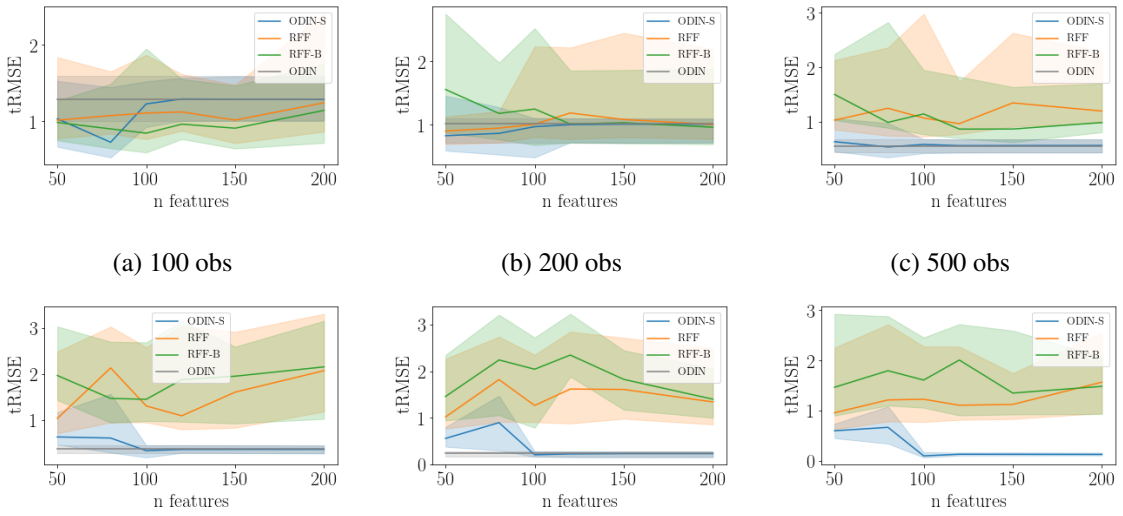
Figure 19: tRMSE vs features for the Lorenz system with noise created using a signal-to-noise ratio of 5.



(a) 100 obs (b) 200 obs (c) 500 obs

(d) 1000 obs (e) 2000 obs (f) 5000 obs

Figure 20: tRMSE vs features for the Lorenz system with noise created using a signal-to-noise ratio of 10.



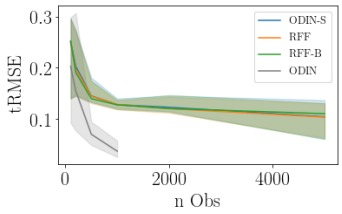
(a) 100 obs (b) 200 obs (c) 500 obs

(d) 1000 obs (e) 2000 obs (f) 5000 obs

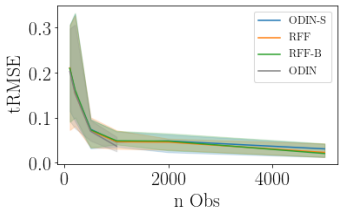
Figure 21: tRMSE vs features for the Lorenz system with noise created using a signal-to-noise ratio of 100.

G.2 Learning Curves

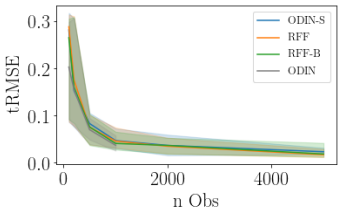
G.2.1 LOTKA VOLTERRA



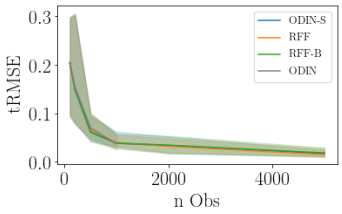
(a) 20 features



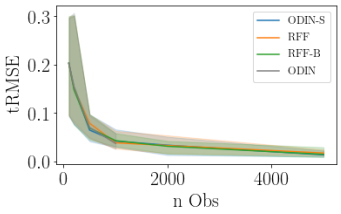
(b) 25 features



(c) 30 features

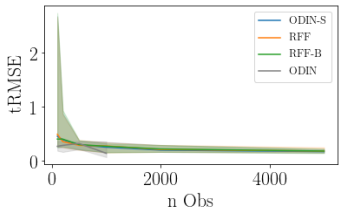


(d) 35 features

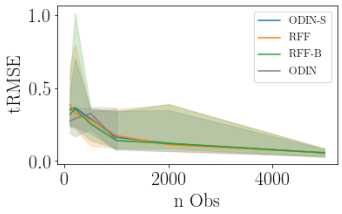


(e) 40 features

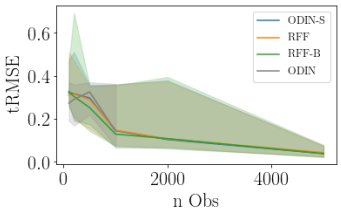
Figure 22: tRMSE vs amount of observations for the Lotka Volterra system with additive noise with $\sigma^2 = 0.1$.



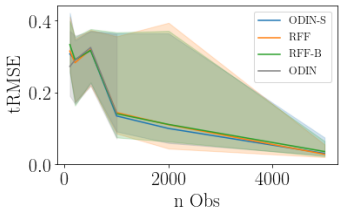
(a) 20 features



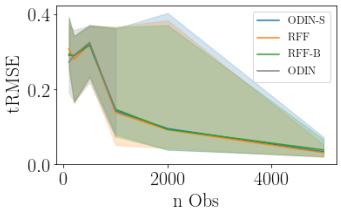
(b) 25 features



(c) 30 features



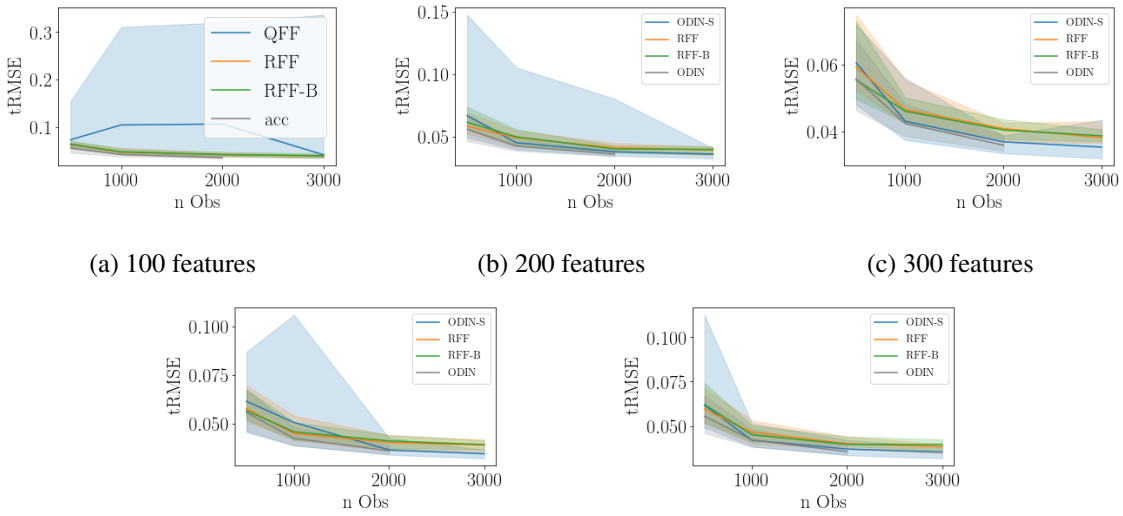
(d) 30 features



(e) 35 features

Figure 23: tRMSE vs amount of observations for the Lotka Volterra system with additive noise with $\sigma^2 = 0.5$.

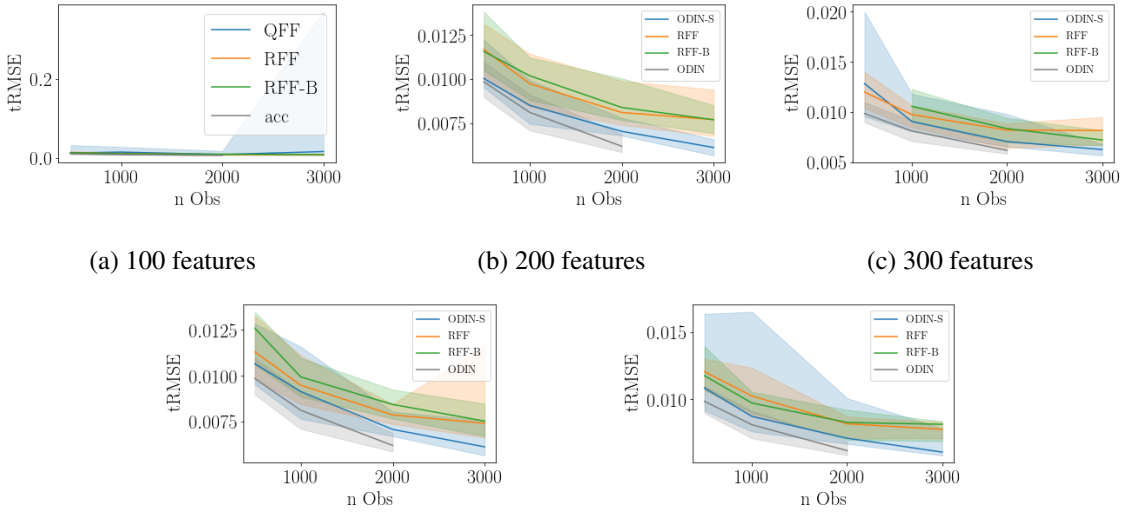
G.2.2 PROTEIN TRANSDUCTION



(a) 100 features (b) 200 features (c) 300 features

(d) 400 features (e) 500 features

Figure 24: tRMSE vs amount of observations for the Protein Transduction system using additive Gaussian noise with $\sigma^2 = 0.01$.

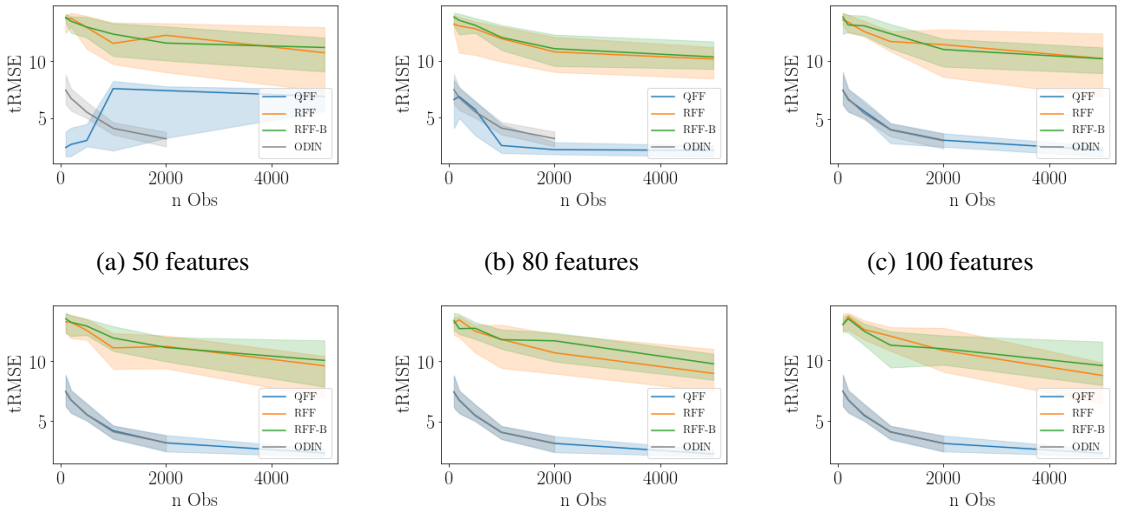


(a) 100 features (b) 200 features (c) 300 features

(d) 400 features (e) 500 features

Figure 25: tRMSE vs amount of observations for the Protein Transduction system using additive Gaussian noise with $\sigma^2 = 0.0001$.

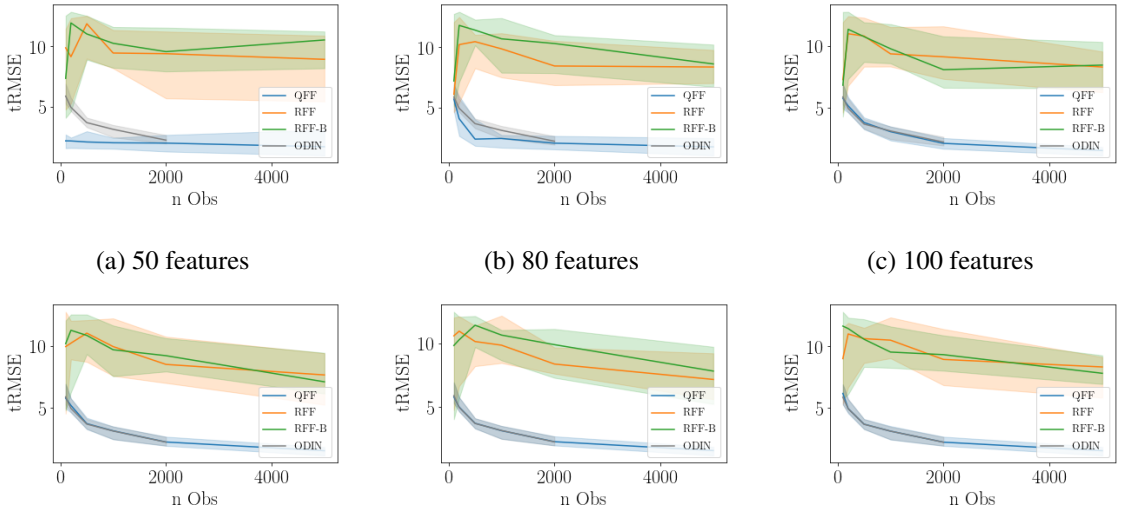
G.2.3 LORENZ



(a) 50 features (b) 80 features (c) 100 features

(d) 120 features (e) 150 features (f) 200 features

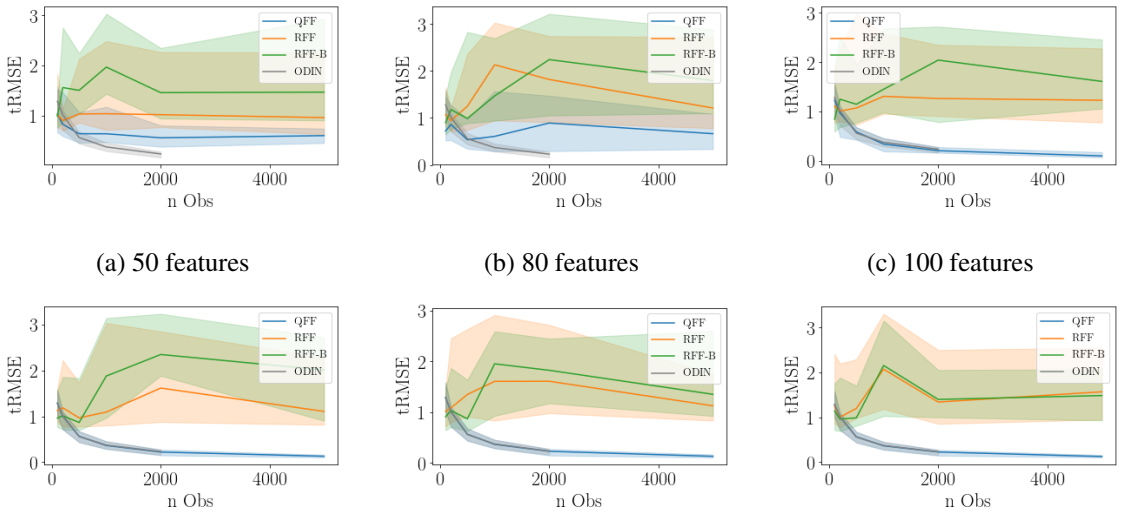
Figure 26: tRMSE vs amount of observations for the Lorenz system with noise created using a signal-to-noise ratio of 5.



(a) 50 features (b) 80 features (c) 100 features

(d) 120 features (e) 150 features (f) 200 features

Figure 27: tRMSE vs amount of observations for the Lorenz system with noise created using a signal-to-noise ratio of 10.

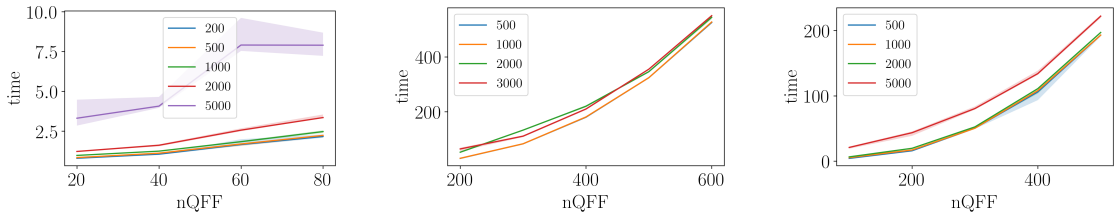


(a) 50 features (b) 80 features (c) 100 features

(d) 120 features (e) 150 features (f) 200 features

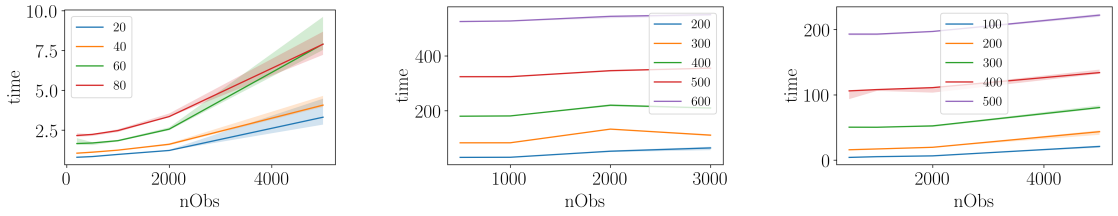
Figure 28: tRMSE vs amount of observations for the Lorenz system with noise created using a signal-to-noise ratio of 100.

G.2.4 RUN TIME



(a) LV (b) PT (c) Lorenz

Figure 29: Run time per iteration in ms vs amount of features for different amounts of observations. As expected from theoretical analysis, the run time per iteration scales approximately cubic.



(a) LV (b) PT (c) Lorenz

Figure 30: Run time per iteration in ms vs amount of observations for different amounts of Fourier features. As expected from theoretical analysis, the run time per iteration scales approximately linear, even though there is a strong bias term.

G.3 Quadrocopter State Inference

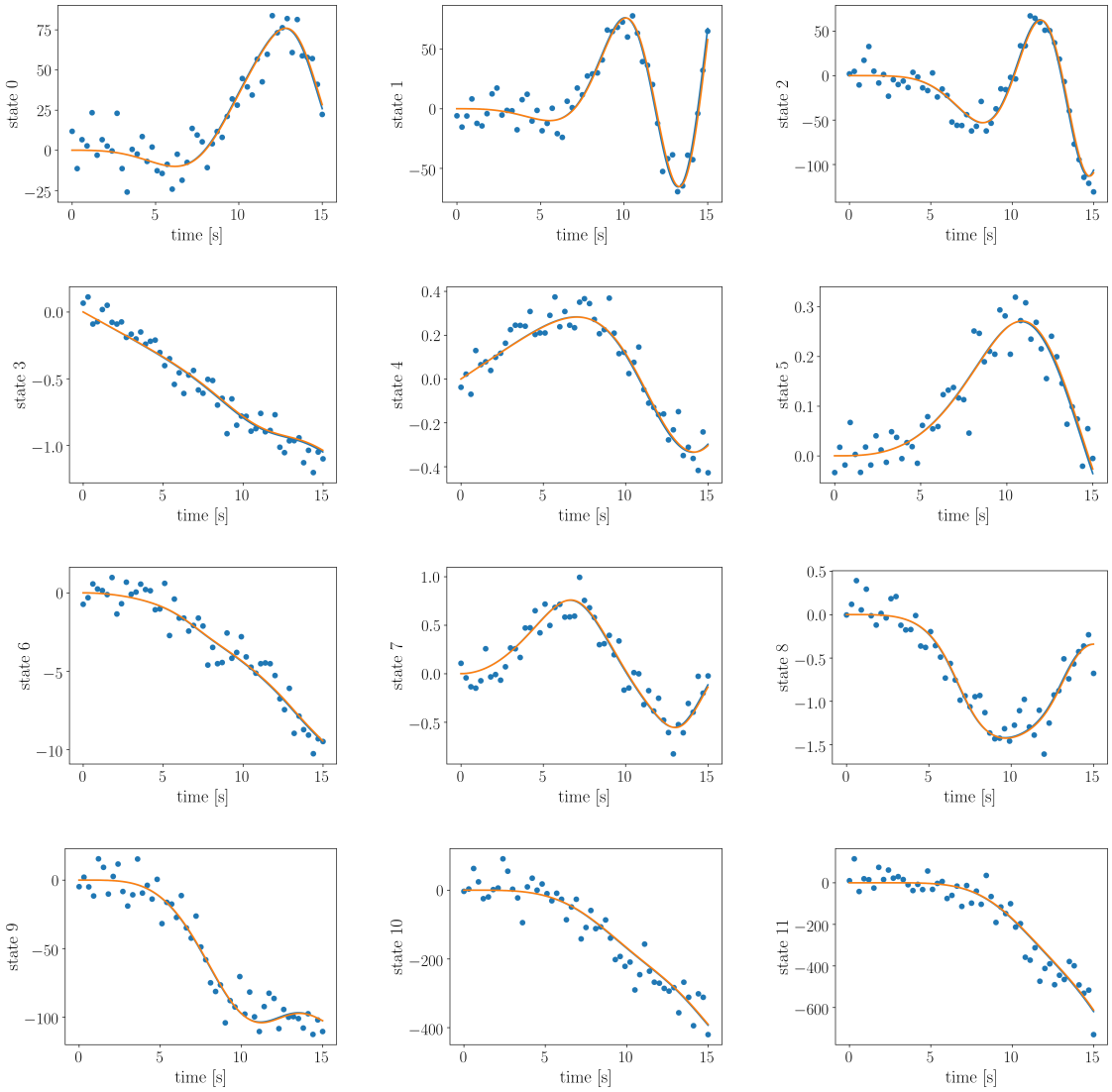


Figure 31: State trajectories obtained by integrating the parameters inferred by ODIN-S (orange). The blue line represents the ground truth, while the blue dots show every 300-th observation for a signal-to-noise ratio of 10.