Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **518,** 5536–5551 (2023) https://doi.org/10.1093/mnras/stac3461
Advance Access publication 2022 November 25

# Radiative transfer as a Bayesian linear regression problem

F. De Ceuster [1,2]★ T. Ceulemans,[1] J. Cockayne,[3] L. Decin[1,4] and J. Yates[2]

[1]*Institute of Astronomy, KU Leuven, Celestijnenlaan 200D, B-3001 Leuven, Belgium*
[2]*Department of Physics and Astronomy, University College London, Gower Place, London WC1E 6BT, UK*
[3]*Department of Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, UK*
[4]*School of Chemistry, University of Leeds, Leeds LS2 9JT, UK*

Accepted 2022 November 21. Received 2022 October 28; in original form 2022 March 9

## ABSTRACT

Electromagnetic radiation plays a crucial role in various physical and chemical processes. Hence, almost all astrophysical simulations require some form of radiative transfer model. Despite many innovations in radiative transfer algorithms and their implementation, realistic radiative transfer models remain very computationally expensive, such that one often has to resort to approximate descriptions. The complexity of these models makes it difficult to assess the validity of any approximation and to quantify uncertainties on the model results. This impedes scientific rigour, in particular, when comparing models to observations, or when using their results as input for other models. We present a probabilistic numerical approach to address these issues by treating radiative transfer as a Bayesian linear regression problem. This allows us to model uncertainties on the input and output of the model with the variances of the associated probability distributions. Furthermore, this approach naturally allows us to create reduced-order radiative transfer models with a quantifiable accuracy. These are approximate solutions to exact radiative transfer models, in contrast to the exact solutions to approximate models that are often used. As a first demonstration, we derive a probabilistic version of the method of characteristics, a commonly-used technique to solve radiative transfer problems.

**Key words:** radiative transfer – methods: numerical – methods: statistical.

## 1 INTRODUCTION

Light, or electromagnetic radiation in general, is a key component of our Universe. Not only does it dictate what we can or cannot observe, it also has the ability to significantly affect numerous physical and chemical processes ranging from radiative heating, cooling, and pressure in hydrodynamics to various photo-reactions in chemistry. As a result, almost every astrophysical simulation requires some form of radiative transfer model.

Over the years, many different schemes have been devised to model radiative transfer, ranging from probabilistic Monte Carlo methods (see e.g. Noebauer & Sim 2019, and the references therein), to several types of formal solvers (see e.g. Kanschat et al. 2009; De Ceuster et al. 2019; and the references therein). Despite many improvements in computational efficiency and the use of modern computer hardware, realistic radiative transfer models keep posing a formidable computational challenge. Consequently, one often has to resort to approximate descriptions of the radiation field, such as flux-limited diffusion (see e.g. Moens et al. 2022), or parametrized radiative heating and cooling functions (see e.g. Xia et al. 2018), which are often used in hydrodynamics models, or semi-analytical descriptions of the photon flux, which are used in photo-chemistry modelling (see e.g. Van de Sande & Millar 2019). Each of these approximate descriptions has its underlying assumptions and limitations, and as models become larger and more complex, it becomes increasingly difficult to properly assess their validity, or to gauge the potential impact of a certain approximation on the results.

Every approximation induces uncertainty on the model result. These uncertainties can either be intrinsic to the model, for instance, due to the discretization of a continuous variable, or are due to the propagation of uncertainties through the model, for instance, due to uncertainties in the radiative constants of the medium that are used in the model. Currently, most radiative transfer models lack any form of uncertainty quantification. This is a severe shortcoming that impedes scientific rigour, in particular, when comparing these seemingly exact model results to naturally noisy observations. In addition, as observations reach ever higher spatial and spectral resolutions (see e.g. Decin et al. 2020), the model uncertainties become ever more relevant. Moreover, with the dawn of a new era of emulated models (see e.g. de Mijolla et al. 2019; Holdship et al. 2021; Kasim et al. 2022), in which algorithms are trained rather than programmed, and simulation data is used as *ground truth*, it is crucial, more than ever, to properly understand the uncertainties associated with simulations. Unfortunately, due to the computational complexity of radiative transfer and the requirement for many approximations it remains very challenging to provide proper uncertainty quantification for most astrophysical radiative transfer models.

There is, however, an approach that can answer both to the need for approximation, because of computational efficiency, and the need for uncertainty quantification, for scientific rigour. Instead of starting from an approximate description, the idea is to start from a more complete description and approximate (or compress) it into a smaller, more tractable, model. There are two key advantages to this approach:

★ E-mail: frederik.deceuster@kuleuven.be

(i) the approximation can be tailored to the problem at hand, and (ii) the uncertainty induced by the approximation can be estimated by the information lost in compressing the model.

In De Ceuster et al. (2020), it was already shown that typical radiative transfer simulations of 3D hydrodynamics models, can often be compressed by more than an order of magnitude in size, without significant loss of accuracy, using a heuristic re-meshing algorithm. This shows that accurate radiative transfer approximations can be obtained by compressing more precise models. It remains, however, to quantify the uncertainties induced by compressing the model and to have more rigorous means to guide the now only heuristic compression algorithm.

In this paper, we propose a novel approach to quantify uncertainties, based on ideas from *probabilistic numerics*. Specifically, we introduce a new numerical method, referred to in the literature as a probabilistic numerical method (Hennig, Osborne & Girolami 2015; Hennig, Osborne & Kersting 2022), whose output is a probability distribution over solutions to the problem. The mean of this distribution coincides with a traditional solution, while the variance can be interpreted as an uncertainty or error measure. The advantages of this probabilistic approach over existing methods are that: (i) since the output contains an intrinsic description of the approximation error, that error can be controlled without the need to compute expensive and often conservative error measures, (ii) as a full probability distribution, the description of error is richer than standard error measures, which typically constitute worst-case bounds on a global (i.e. norm-wise) or local error, and (iii) since the approximation error is expressed in a probabilistic manner, it can naturally be combined with other sources of uncertainty to provide a unified description of all uncertainty in the solution, as will be demonstrated in Section 3.2.2. These three points make probabilistic numerical methods particularly appropriate for modelling radiative transfer, given the dire need for reliable uncertainty estimates on a computationally expensive model in the presence of uncertainties on input quantities such as radiative constants.

In particular, we propose to treat radiative transfer as a Bayesian linear regression problem. The radiation field is modelled as the expectation of a multivariate Gaussian probability distribution over possible solutions for the radiation field, conditioned on evaluations of the radiative transfer equation and boundary conditions. As such, the variance of the conditioned distribution can be used as a measure of uncertainty on this result. The computational complexity of the regression model can be controlled by the dimension of the feature space, i.e. the number of basis functions that is used. This allows us to create approximate (reduced-order) radiative transfer models, by reducing the number of basis functions.

The idea to treat function approximation or the solution of operator equations as a regression problem is certainly not new and can be traced all the way back to Poincaré (1896), as pointed out by Diaconis (1988). More recently, motivated by Hennig et al. (2015), these ideas gained renewed interest and now form an active research domain in applied mathematics known as probabilistic numerics (see e.g. Cockayne et al. 2019; Hennig et al. 2022). For a comprehensive recent history, see e.g. Oates & Sullivan (2019). The specific idea to view the solution of operator equations as a Bayesian linear regression problem has been proposed several times, by several different authors, and in several different contexts (e.g. van den Boogaart 2001; Graepel 2003; Cockayne et al. 2017). While, in the literature, this is usually derived from a Gaussian process, here, we take a slightly more general point of view.

The probabilistic numerical method presented here is closely related to finite element methods (see also e.g. Girolami et al. 2021). In fact, the method just gives a probabilistic interpretation to an otherwise classical collocation method (see e.g. Kansa 1990a,b; Fasshauer 1999). Finite element methods were introduced in the context of astrophysical radiative transfer by Dykema, Klein & Castor (1996), who applied it to the moments of the radiative transfer equation. Since then, these methods have successfully been applied in several astrophysical contexts (see e.g. Meier 1999; Richling et al. 2001; Korčáková & Kubát 2003). Due to their widespread use, especially in industry, there is a vast body of research dedicated to uncertainty quantification for these methods (see e.g. Verfürth 2013, and the references therein). Also in the astrophysical context, for instance, Richling et al. (2001) proposed an error measure on their finite element radiative transfer solver, which they furthermore used to adapt their discretization. The key difference between classical finite element methods and the method presented here, is the probabilistic interpretation of the results, i.e. here the solutions are (conditioned) probability distributions over the space of possible solutions, rather than a single solution. This allows us to also take into account the uncertainties on the model input, and furthermore facilitates the use of our model both in forward and inverse modelling pipelines.

Alternatively, the method presented in this paper can be viewed as a linear (and hence analytically solvable) version of a physics-informed neural network method (see e.g. Lagaris, Likas & Fotiadis 1998; Lagaris, Likas & Papageorgiou 2000; Raissi, Perdikaris & Karniadakis 2019). This technique, inspired by machine learning, to solve, for instance operator equations, has already been successfully applied to radiative transfer problems, for example by Mishra & Molinaro (2021), who, furthermore, derived rigorous error bounds for their results (Mishra & Molinaro 2022). The assumption of linearity makes our model much simpler than this, and allows us to obtain analytic solutions which can be used, for instance, to relate it directly to the commonly-used method of characteristics.

The structure of this paper is as follows. In Section 2, we introduce Bayesian linear regression and show how it can be used to solve linear operator equations in a way that naturally allows for uncertainty quantification. In Section 3, we apply this to radiative transfer. We show how reduced-order radiative transfer models can be obtained, and we derive a Bayesian version of the method of characteristics. Section 4 concerns future research towards practical implementations, and we conclude with Section 5.

## 2 METHODS

We present a probabilistic numerical method to solve linear operator equations by treating it as a (Bayesian) linear regression problem. This idea has already been discussed at length in the literature (see e.g. van den Boogaart 2001; Graepel 2003; Cockayne et al. 2017). Nevertheless, we present it again, but in a slightly more general way, to demonstrate its full potential for astrophysical modelling, and radiative transfer in particular. For a more comprehensive introduction, see e.g. Bishop (2006), Rasmussen & Williams (2006), or Hennig et al. (2022).

### 2.1 Linear regression

The aim of a linear regression model is to approximate (or fit) a function, $f$, with a linear combination of basis functions, $\phi_i$, based on data in the form of function evaluations, $(x_d, y_d \equiv f(x_d))$. In this paper, we only consider real functions, so all variables are always assumed to be real. Given a set of $N_\mathrm{b}$ basis functions, $\{\phi_i\}$, and a set of $N_\mathrm{d}$ data points, $\{(x_d, y_d)\}$, the approximation can either be expanded in terms

of the basis function or in terms of the data, resulting respectively in the primal and dual formulation.

### 2.1.1 Primal formulation

In the primal formulation, the approximation, $\tilde{f}(x)$, is modelled as a linear combination of the basis functions,

$$\tilde{f}(x) \;=\; \sum_{i=1}^{N_\mathrm{b}} w_i\, \phi_i(x) \;\equiv\; \boldsymbol{w}^\mathrm{T}\boldsymbol{\phi}(x), \tag{1}$$

where we defined the weight vector, $\boldsymbol{w}$, and basis function vector, $\boldsymbol{\phi}$. Appropriate weights, $w_i$, can be found, for instance, by minimizing the regularized mean squared error between the model and the data,

$$\mathrm{RMSE}(\boldsymbol{w}) \;\equiv\; \textstyle\sum_{d=1}^{N_\mathrm{d}} \frac{1}{\sigma_d^2} \left(\boldsymbol{w}^\mathrm{T}\boldsymbol{\phi}(x_d) - y_d\right)^2 \;+\; \sum_{i=1}^{N_\mathrm{b}} \left(\frac{w_i}{\lambda_i}\right)^2. \tag{2}$$

The factors, $\sigma_d^{-2}$, weight the contributions of the different data points to the mean error, and are summarized in the diagonal matrix $\boldsymbol{\sigma} \equiv \mathrm{diag}(\sigma_d)$. We also added a regularization term, characterized by the diagonal matrix $\boldsymbol{\lambda} \equiv \mathrm{diag}(\lambda_i)$, which penalizes the size of the components of the weight vectors. This will guarantee the existence of a unique solution, as we will see below. If we define the design matrix, $\Phi_{di} \equiv \phi_i(x_d)$, and the data vector pair, $(\boldsymbol{x}, \boldsymbol{y})$, equation (2) can conveniently be rewritten as

$$\mathrm{RMSE}(\boldsymbol{w}) \;\equiv\; \left(\boldsymbol{\sigma}^{-1}\left(\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{y}\right)\right)^2 + \left(\boldsymbol{\lambda}^{-1}\boldsymbol{w}\right)^2, \tag{3}$$

in which the square of a vector, $\boldsymbol{a}$, is defined as $(\boldsymbol{a})^2 \equiv \boldsymbol{a}^\mathrm{T}\boldsymbol{a}$. Minimizing this regularized mean squared error by demanding a vanishing gradient with respect to the weights, $\boldsymbol{w}$, yields

$$\left(\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi} + \boldsymbol{\lambda}^{-2}\right)\boldsymbol{w}_\mathrm{min} \;=\; \boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\sigma}^{-2}\boldsymbol{y}, \tag{4}$$

in which $\boldsymbol{w}_\mathrm{min}$ is the weight vector that minimizes (3). The resulting (optimal) function approximation (1) is thus given by

$$\tilde{f}(x) \;=\; \boldsymbol{y}^\mathrm{T}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi} + \boldsymbol{\lambda}^{-2}\right)^{-1}\boldsymbol{\phi}(x). \tag{5}$$

The inverse is guaranteed to exist as long as the regularization term is non-zero, i.e. $\lambda_i \neq 0$, $\forall i \in \{1, \ldots, N_\mathrm{b}\}$. Note that a $(N_\mathrm{b} \times N_\mathrm{b})$-dimensional linear system must be solved to obtain the approximate solution, and hence the computational cost of the primal formulation is determined by the number of basis functions, $N_\mathrm{b}$.

### 2.1.2 Dual formulation

In the dual formulation, the approximation, $\tilde{f}(x)$, is modelled as a linear combination of (evaluations of a kernel function on) the data,

$$\tilde{f}(x) \;=\; \sum_{d=1}^{N_\mathrm{d}} v_d\, k(x_d, x) \;\equiv\; \boldsymbol{v}^\mathrm{T}k(\boldsymbol{x}, x), \tag{6}$$

in which the kernel is defined in terms of the basis functions,

$$k(x, x') \;\equiv\; \sum_{i=1}^{N_\mathrm{b}} \phi_i(x)\lambda_i^2\phi_i(x') \;=\; \boldsymbol{\phi}(x)^\mathrm{T}\boldsymbol{\lambda}^2\boldsymbol{\phi}(x'), \tag{7}$$

where we used the regularization parameter, $\boldsymbol{\lambda}$, from equation (2). This definition of the kernel ensures the correspondence between the primal and dual formulation (see also Section 2.1.3). Intuitively, the kernel expresses how the solution hinges on the data points, $\boldsymbol{x}$. From this definition of the kernel, it can be seen that the weights of the primal and dual formulation are related as $\boldsymbol{w} = \boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{v}$. Also here, the appropriate weights can be obtained by minimizing the regularized mean squared error. Keeping our previous definitions, the error (3) in terms of the new weights, $\boldsymbol{v}$, reads

$$\mathrm{RMSE}(\boldsymbol{v}) \;\equiv\; \left(\boldsymbol{\sigma}^{-1}\left(\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{v} - \boldsymbol{y}\right)\right)^2 + \left(\boldsymbol{\lambda}\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{v}\right)^2. \tag{8}$$

Minimizing this regularized mean squared error by demanding a vanishing gradient with respect to the new weights, $\boldsymbol{v}$, yields

$$\left(\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\sigma}^{-2}\,\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T} \;+\; \boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}\right)\boldsymbol{v}_\mathrm{min} \;=\; \boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\sigma}^{-2}\,\boldsymbol{y}, \tag{9}$$

in which $\boldsymbol{v}_\mathrm{min}$ is the weight vector that minimizes (8). Note that $\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}$ might not be invertible and thus equation (9) might not have a unique solution. However, we can always pick the uniquely solvable system that will also minimize (8), by omitting the overall factor, $\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\sigma}^{-2}$, which yields

$$\left(\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T} \;+\; \boldsymbol{\sigma}^2\right)\boldsymbol{v}_\mathrm{min} \;=\; \boldsymbol{y}. \tag{10}$$

The resulting function approximation then reads

$$\tilde{f}(x) \;=\; \boldsymbol{y}^\mathrm{T}\left(\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T} \;+\; \boldsymbol{\sigma}^2\right)^{-1}k(\boldsymbol{x}, x). \tag{11}$$

Note that the inverse is guaranteed to exist as long as $\sigma_i \neq 0$, $\forall i \in \{1, \ldots, N_\mathrm{d}\}$. In this case, a $(N_\mathrm{d} \times N_\mathrm{d})$-dimensional linear system needs to be solved to obtain the approximate solution, and thus, in contrast to the primal formulation, the computational cost of the dual formulation is determined by the number of data points, $N_\mathrm{d}$.

Note that the dual formulation can also be constructed directly from a given kernel without any link to a set of basis functions. In particular, the design matrix always appears as, $\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T} = k(\boldsymbol{x}, \boldsymbol{x})$, and thus can always be replaced by its equivalent kernel expression.

### 2.1.3 Primal versus dual formulation

One can show that the solutions of the primal (5) and dual (11) formulation are equal. For this to be true, one needs to show that

$$\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi} + \boldsymbol{\lambda}^{-2}\right)^{-1} \;=\; \left(\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T} + \boldsymbol{\sigma}^2\right)^{-1}\boldsymbol{\Phi}\boldsymbol{\lambda}^2, \tag{12}$$

as is done in Appendix A1. The only, yet key, difference between both formulations is thus the size of the linear system that needs to be solved to obtain the solution.

### 2.1.4 Solving linear partial differential equations as linear regression

Numerically solving linear operator equations, and in particular linear partial differential equations (PDEs), can be viewed as a linear regression problem. Say we want to numerically solve a PDE,

$$\begin{aligned} &\mathscr{L}f(x) = g(x), \quad x \in D \\ &\mathscr{B}f(x) = h(x), \quad x \in \partial D \end{aligned} \tag{13}$$

on a domain, $D$, with boundary, $\partial D$, where the PDE and boundary conditions are determined respectively by the linear operators $\mathscr{L}$ and $\mathscr{B}$. Suppose that for the numerical solution the domain is discretized to $\tilde{D}$, and that $\boldsymbol{a}$ is a vector containing the points in $\tilde{D}$, and the boundary is discretized to $\partial\tilde{D}$, and that $\boldsymbol{b}$ is a vector containing the points in $\partial\tilde{D}$, then we can split the data as,

$$(\boldsymbol{x}, \boldsymbol{y}) \;=\; \left(\begin{pmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{pmatrix}, \begin{pmatrix}g(\boldsymbol{a})\\ h(\boldsymbol{b})\end{pmatrix}\right) \;\equiv\; \left(\begin{pmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{pmatrix}, \begin{pmatrix}\boldsymbol{g}\\ \boldsymbol{h}\end{pmatrix}\right). \tag{14}$$

Similarly, we can split the matrix, $\boldsymbol{\sigma}$, as, $\boldsymbol{\sigma} = \mathrm{diag}(\boldsymbol{\sigma}_\mathrm{L}, \boldsymbol{\sigma}_\mathrm{B})$, and the design matrix, $\boldsymbol{\Phi}$, which has function evaluations at different data points on its rows, can be split as,

$$\boldsymbol{\Phi} \;\equiv\; \begin{pmatrix}\mathscr{L}\boldsymbol{\phi}(\boldsymbol{a})\\ \mathscr{B}\boldsymbol{\phi}(\boldsymbol{b})\end{pmatrix} \;\equiv\; \begin{pmatrix}\boldsymbol{\Phi}_\mathrm{L}\\ \boldsymbol{\Phi}_\mathrm{B}\end{pmatrix}. \tag{15}$$

In this new notation, the key matrices appearing in the primal and dual formulation can respectively be written as

$$\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi} + \boldsymbol{\lambda}^{-2} = \boldsymbol{\Phi}_{\mathrm{L}}^{\mathrm{T}}\boldsymbol{\sigma}_{\mathrm{L}}^{-2}\boldsymbol{\Phi}_{\mathrm{L}} + \boldsymbol{\Phi}_{\mathrm{B}}^{\mathrm{T}}\boldsymbol{\sigma}_{\mathrm{B}}^{-2}\boldsymbol{\Phi}_{\mathrm{B}} + \boldsymbol{\lambda}^{-2} \tag{16}$$

$$\boldsymbol{\Phi}\boldsymbol{\lambda}^{2}\boldsymbol{\Phi}^{\mathrm{T}} + \boldsymbol{\sigma}^{2} = \begin{pmatrix} \boldsymbol{\Phi}_{\mathrm{L}}\boldsymbol{\lambda}^{2}\boldsymbol{\Phi}_{\mathrm{L}}^{\mathrm{T}} + \boldsymbol{\sigma}_{\mathrm{L}}^{2} & \boldsymbol{\Phi}_{\mathrm{L}}\boldsymbol{\lambda}^{2}\boldsymbol{\Phi}_{\mathrm{B}}^{\mathrm{T}} \\ \boldsymbol{\Phi}_{\mathrm{B}}\boldsymbol{\lambda}^{2}\boldsymbol{\Phi}_{\mathrm{L}}^{\mathrm{T}} & \boldsymbol{\Phi}_{\mathrm{B}}\boldsymbol{\lambda}^{2}\boldsymbol{\Phi}_{\mathrm{B}}^{\mathrm{T}} + \boldsymbol{\sigma}_{\mathrm{L}}^{2} \end{pmatrix}. \tag{17}$$

In terms of the kernel, equation (17) can also be rewritten as

$$k(\boldsymbol{x},\boldsymbol{x}) + \boldsymbol{\sigma}^{2} = \begin{pmatrix} \mathscr{L}_1\mathscr{L}_2 k(\boldsymbol{a},\boldsymbol{a}) + \boldsymbol{\sigma}_{\mathrm{L}}^{2} & \mathscr{L}_1\mathscr{B}_2 k(\boldsymbol{a},\boldsymbol{b}) \\ \mathscr{B}_1\mathscr{L}_2 k(\boldsymbol{b},\boldsymbol{a}) & \mathscr{B}_1\mathscr{B}_2 k(\boldsymbol{b},\boldsymbol{b}) + \boldsymbol{\sigma}_{\mathrm{B}}^{2} \end{pmatrix}, \tag{18}$$

in which the subscripts on the operators indicate whether they act on the first or second argument of the kernel.

As with linear regression, numerically solving the PDE can thus be formulated as a minimization problem and can be solved both in the primal and dual formulation. The only difference is that the design matrix, $\boldsymbol{\Phi}$, should be redefined as in equation (15). This technique is known as the collocation method for solving operator equations (see e.g. Fasshauer 1999; Schaback & Wendland 2006).

Intuitively, this can be understood as follows. The weights for the solution of the linear operator equation (13) are determined in terms of the basis functions, $\{\phi_i\}$, by fitting the functions $g(x)$ and $h(x)$, with the basis functions $\{\mathscr{L}\phi_i\}$ and $\{\mathscr{B}\phi_i\}$ respectively. Hence, the basis functions should ideally be chosen such that $\{\mathscr{L}\phi_i\}$ can properly fit $g(x)$, $\{\mathscr{B}\phi_i\}$ can properly fit $h(x)$, and $\{\phi_i\}$ can properly fit the sought after solution function $f(x)$.

### 2.2 Bayesian linear regression

The framework of linear regression can also be derived in a Bayesian or probabilistic setting. Here we consider a stochastic function, $F(x)$, giving a probability distribution over the possible results for every value, $x$, and are interested in the distribution of this function as it is conditioned on observations of evaluations $(\boldsymbol{x}, \boldsymbol{y})$ of that function, i.e. our goal is to find $p(F(x)\,|\,\boldsymbol{y})$. Note that, to simplify notation further on, we write $|\boldsymbol{y}$ to denote conditioning on the data, whereas we actually mean $|(\boldsymbol{x}, \boldsymbol{y})$. We summarized the definitions and some further explanations of the statistical concepts that are used throughout this and subsequent sections in Appendix A2.

#### 2.2.1 *Bayesian primal formulation*

Given a linear model in the primal formulation with corresponding weights, $\boldsymbol{w}$, a zero-mean Gaussian error on the observed function evaluations, $\boldsymbol{\mathcal{Y}}$, results in a Gaussian likelihood given by

$$p\left(\boldsymbol{\mathcal{Y}}\,|\,\boldsymbol{w}\right) = \mathcal{N}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}),\, \boldsymbol{\sigma}^{2}\right) = \mathcal{N}\left(\boldsymbol{\Phi}\boldsymbol{w},\, \boldsymbol{\sigma}^{2}\right). \tag{19}$$

Note that we reused the variable $\boldsymbol{\sigma}^2$ and reinterpreted it as the variance on the data, allowing for deviations from the mean value, $\boldsymbol{\Phi}\boldsymbol{w}$, predicted by the model given the weights, $\boldsymbol{w}$. We will see below that both interpretations are indeed compatible. Furthermore, we assume a zero-mean Gaussian prior on the stochastic weights, $\boldsymbol{\mathcal{W}}$,

$$p\left(\boldsymbol{\mathcal{W}}\right) = \mathcal{N}\left(\boldsymbol{0},\, \boldsymbol{\lambda}^{2}\right). \tag{20}$$

Note that we reused the variable $\boldsymbol{\lambda}^2$ and reinterpreted it as the variance of the prior on the weights. Using the relations given in Appendix A3, we can infer that the implied distribution of the weights conditioned on the data is given by

$$p(\boldsymbol{\mathcal{W}}\,|\,\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{y}},\, \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{y}}\right), \tag{21}$$

in which the mean vector and covariance matrix are defined as

$$\boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{y}} \equiv \boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{y}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\boldsymbol{y}, \tag{22}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{y}} \equiv \left(\boldsymbol{\lambda}^{-2} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}\right)^{-1}. \tag{23}$$

Since the stochastic function, $F(x)$, is a linear mapping of the weights, $F(x) = \boldsymbol{\mathcal{W}}^{\mathrm{T}}\boldsymbol{\phi}(x)$, the conditioned distribution reads

$$p\left(F(x)\,|\,\boldsymbol{y}\right) = \mathcal{N}\left(\mu_{\mathrm{primal}}(x),\, \sigma^{2}_{\mathrm{primal}}(x)\right), \tag{24}$$

in which the mean and variance are defined as

$$\mu_{\mathrm{primal}}(x) \equiv \mu_{\boldsymbol{w}|\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{\phi}(x), \tag{25}$$

$$\sigma^{2}_{\mathrm{primal}}(x) \equiv \boldsymbol{\phi}(x)^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{w}|\boldsymbol{y}}\boldsymbol{\phi}(x). \tag{26}$$

Substituting equation (22), we rediscover the primal solution (5) as the mean of the resulting conditioned primal distribution,

$$\mu_{\mathrm{primal}}(x) = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi} + \boldsymbol{\lambda}^{-2}\right)^{-1}\boldsymbol{\phi}(x). \tag{27}$$

Furthermore, we now also have a measure for the spread in possible approximations from the variance of the conditioned distribution,

$$\sigma^{2}_{\mathrm{primal}}(x) = \boldsymbol{\phi}(x)^{\mathrm{T}}\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi} + \boldsymbol{\lambda}^{-2}\right)^{-1}\boldsymbol{\phi}(x). \tag{28}$$

This allows us to predict an approximation for the function, $f$, based on the data, $(\boldsymbol{x}, \boldsymbol{y})$, and provide a confidence level for the result. It should be noted that to compute the variance either for each $x$ a separate $(N_{\mathrm{b}} \times N_{\mathrm{b}})$-dimensional linear system needs to be solved, or that an $(N_{\mathrm{b}} \times N_{\mathrm{b}})$-dimensional matrix needs to be inverted explicitly. However, since one usually does not require a high precision for an uncertainty estimate, the matrix inverse can quickly be computed in an approximate way.

#### 2.2.2 *Bayesian dual formulation*

A similar argument can be made for the dual formulation and is typically encountered in the context of Gaussian processes (see e.g. Rasmussen & Williams 2006). Since we assumed that the weights, $\boldsymbol{\mathcal{W}}$, and the errors (or spread) in the data, $\boldsymbol{\mathcal{Y}}$, both follow a zero-mean (multivariate) Gaussian distribution, the function values and the data will follow a joint (multivariate) Gaussian distribution,

$$p\left(\begin{bmatrix} F(x) \\ \boldsymbol{y} \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} 0 \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} k(x,x) & k(x,\boldsymbol{x}) \\ k(\boldsymbol{x},x) & k(\boldsymbol{x},\boldsymbol{x}) + \boldsymbol{\sigma}^{2} \end{bmatrix}\right), \tag{29}$$

where we reused the definition of the kernel (7). The posterior distribution can be obtained by conditioning on the data $(\boldsymbol{x}, \boldsymbol{y})$, using the relations given in Appendix A4,

$$p\left(F(x)\,|\,\boldsymbol{y}\right) = \mathcal{N}\left(\mu_{\mathrm{dual}}(x),\, \sigma^{2}_{\mathrm{dual}}(x)\right), \tag{30}$$

in which the mean and variance are respectively defined as

$$\mu_{\mathrm{dual}}(x) \equiv k(x,\boldsymbol{x})\left(k(\boldsymbol{x},\boldsymbol{x}) + \boldsymbol{\sigma}^{2}\right)^{-1}\boldsymbol{y} \tag{31}$$

$$\sigma^{2}_{\mathrm{dual}}(x) \equiv k(x,x) - k(x,\boldsymbol{x})\left(k(\boldsymbol{x},\boldsymbol{x}) + \boldsymbol{\sigma}^{2}\right)^{-1}k(\boldsymbol{x},x). \tag{32}$$

Rewriting this in terms of the design matrix, we rediscover the dual solution (11) as expectation of the conditioned distribution,

$$\mu_{\mathrm{dual}}(x) = \boldsymbol{\phi}(x)^{\mathrm{T}}\boldsymbol{\lambda}^{2}\boldsymbol{\Phi}^{\mathrm{T}}\left(\boldsymbol{\Phi}\boldsymbol{\lambda}^{2}\boldsymbol{\Phi}^{\mathrm{T}} + \boldsymbol{\sigma}^{2}\right)^{-1}\boldsymbol{y}. \tag{33}$$

Moreover, we can similarly obtain a measure for the quality of the approximation from the variance of the conditioned distribution,

$$\sigma^{2}_{\mathrm{dual}}(x) = \boldsymbol{\phi}(x)^{\mathrm{T}}\left(\boldsymbol{\lambda}^{2} - \boldsymbol{\lambda}^{2}\boldsymbol{\Phi}^{\mathrm{T}}\left(\boldsymbol{\Phi}\boldsymbol{\lambda}^{2}\boldsymbol{\Phi}^{\mathrm{T}} + \boldsymbol{\sigma}^{2}\right)^{-1}\boldsymbol{\Phi}\boldsymbol{\lambda}^{2}\right)\boldsymbol{\phi}(x). \tag{34}$$

Again, we can predict an approximation for the function, $f$, based on the data, $(\boldsymbol{x}, \boldsymbol{y})$, and provide a confidence level for the result. Also here, it should be noted that to compute the variance either for each $x$ a separate $(N_\mathrm{d} \times N_\mathrm{d})$-dimensional linear system needs to be solved, or that an $(N_\mathrm{d} \times N_\mathrm{d})$-dimensional matrix needs to be inverted explicitly. However, as in the primal formulation, since one usually does not require a high precision for an uncertainty estimate, the matrix inverse can quickly be computed in an approximate way.

#### 2.2.3 *Bayesian primal versus Bayesian dual formulation*

As in the non-Bayesian case, we note that in the primal formulation a $(N_\mathrm{b} \times N_\mathrm{b})$-dimensional linear system needs to be solved, while in the dual formulation it is a $(N_\mathrm{d} \times N_\mathrm{d})$-dimensional linear system.

We already showed in the non-Bayesian case (Section 2.1.3) that the primal and dual solutions are equal. Using the Woodburry matrix identity, we can now also easily verify that the variances for the primal and dual Bayesian formulation are equal, since,

$$\left(\boldsymbol{\lambda}^{-2} + \boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}\right)^{-1} = \boldsymbol{\lambda}^2 - \boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}\left(\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T} + \boldsymbol{\sigma}^2\right)^{-1}\boldsymbol{\Phi}\boldsymbol{\lambda}^2. \quad (35)$$

We can conclude that the duality also holds in the probabilistic sense, which implies for the probability distributions that

$$\mathcal{N}\left(\mu_\mathrm{primal}(x),\, \sigma^2_\mathrm{primal}(x)\right) = \mathcal{N}\left(\mu_\mathrm{dual}(x),\, \sigma^2_\mathrm{dual}(x)\right). \quad (36)$$

Both formulations are thus equivalent and can therefore be used interchangeably, as long as they are both well defined.

It should be noted that our choice of Gaussian priors was only motivated by computational convenience, and that it is not ideal. For instance, the Gaussian distribution always assigns a non-zero probability, also to negative values of a variable. For many physical quantities that are only positive, such as density or temperature, this is not desirable as it can lead to non-physical results. However, this is the case for many numerical schemes, and, bearing in mind these dangers, the Gaussian distribution is a good first approximation for the uncertainties in our variables.

The Bayesian linear regression problem can alternatively also be formulated using other distributions for the priors (see e.g. Shah, Wilson & Ghahramani 2014, for an example using Student-$t$ distributed priors), but always at the expense of computational convenience.

#### 2.2.4 *The limit of uninformative data: $\boldsymbol{\sigma} \to \boldsymbol{\infty}$*

In order to gain more insight into these results, we consider some limiting cases. In the limit of uninformative data, i.e. $\boldsymbol{\sigma} \to \boldsymbol{\infty}$, the uncertainty on the data is so large that conditioning on them does not change the prior distribution. Hence, as can be seen by taking the limit, $\boldsymbol{\sigma} \to \boldsymbol{\infty}$, in equations (27, 28, 33, 34), one finds

$$\mu_\mathrm{primal}(x) = \mu_\mathrm{dual}(x) = 0, \quad (37)$$

$$\sigma^2_\mathrm{primal}(x) = \sigma^2_\mathrm{dual}(x) = \boldsymbol{\phi}(x)^\mathrm{T}\boldsymbol{\lambda}^2\boldsymbol{\phi}(x), \quad (38)$$

which is exactly the zero-mean prior distribution that we assumed.

A similar argument[1] can be made for the limit of perfect prior knowledge, i.e. $\boldsymbol{\lambda} \to \mathbf{0}$, when the confidence in the prior is so large that no conditioning on any data can change it.

[1]The reason why a similar argument applies is the duality between the parameters $\boldsymbol{\sigma}$ and $\boldsymbol{\lambda}$. For instance, in the simplified case that $\boldsymbol{\sigma} = \sigma\mathbb{1}$ and $\boldsymbol{\lambda} = \lambda\mathbb{1}$, the parameter determining the behaviour of the model is $\sigma/\lambda$.

#### 2.2.5 *The limit of perfect data: $\boldsymbol{\sigma} \to \mathbf{0}$*

In order to gain further insight into the results, let us ignore any effects that might be caused by uncertainties in the data and consider the limit of perfect data, i.e. $\boldsymbol{\sigma} \to \mathbf{0}$. The primal and dual solutions in this limit are respectively given by

$$\mu_\mathrm{primal}(x) \to \boldsymbol{y}^\mathrm{T}\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\phi}(x), \quad (39)$$

$$\mu_\mathrm{dual}(x) \to \boldsymbol{y}^\mathrm{T}\left(\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}\right)^{-1}\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\phi}(x). \quad (40)$$

Note that the inverses above do not necessarily exist. In particular, if $N_\mathrm{d} > N_\mathrm{b}$, the singular value decomposition shows that $\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}$ must be singular and thus only the primal formulation remains, whereas if $N_\mathrm{d} < N_\mathrm{b}$, it follows that $\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\Phi}$ must be singular and thus only the dual formulation remains.[2] As a result, in the limit $\boldsymbol{\sigma} \to \mathbf{0}$, if $N_\mathrm{d} \neq N_\mathrm{b}$, the duality between the two formulations ceases to hold and only the formulation with the smallest corresponding linear system will have a unique solution.

Moreover, note that in this limit the variance of the primal formulation always vanishes. As a result, there is no probabilistic interpretation in the limit of perfect data when $N_\mathrm{d} > N_\mathrm{b}$ and only the primal formulation remains. Intuitively, this can be understood since, in general, a linear regression model using $N_\mathrm{b}$ basis functions cannot perfectly fit $N_\mathrm{d}$ data points. Hence, the assumption that the data can be fit perfectly, will, in general, be wrong. The uncertainty in the data, i.e. $\boldsymbol{\sigma} \neq \mathbf{0}$, is required to allow for some slack in fitting the $N_\mathrm{d}$ data points with only $N_\mathrm{b}$ basis functions.

Similarly, in the dual formulation, the variance in the limit of perfect data, i.e. $\boldsymbol{\sigma} \to \mathbf{0}$, reads

$$\sigma^2_\mathrm{dual}(x) = \boldsymbol{\phi}(x)^\mathrm{T}\left(\boldsymbol{\lambda}^2 - \boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}\left(\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}\right)^{-1}\boldsymbol{\Phi}\boldsymbol{\lambda}^2\right)\boldsymbol{\phi}(x), \quad (41)$$

which evidently only makes sense if the inverse of $\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}$ exists, which requires that $N_\mathrm{d} \leq N_\mathrm{b}$. In the particular case that $N_\mathrm{d} = N_\mathrm{b}$, demanding that $\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}$ is invertible implies that $\boldsymbol{\Phi}$ is invertible, such that also in this case the variance vanishes. Hence, in the limit of perfect data there is only a probabilistic interpretation if $N_\mathrm{d} < N_\mathrm{b}$, assuming that the inverse for $\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}$ exits. Intuitively this can be understood from the fact that we model the spread in the distribution with the same basis functions as we use to model the function approximation. If we assume the data to be exact and if $N_\mathrm{d} \geq N_\mathrm{b}$, the contributions of all basis functions are fixed by the data and there are no undetermined degrees of freedom that can cause a spread in the resulting distribution conditioned on the data.

A similar argument can be made for the limit of uninformative prior knowledge, i.e. $\boldsymbol{\lambda} \to \boldsymbol{\infty}$, when the uncertainty in the prior is so large that the regression essentially fully depends on the data.

### 2.3 Uncertainty quantification

Quantifying uncertainties is an approximate endeavour. After all, the exact solution, $f$, is required in order to determine the exact error, $\varepsilon$, that is made in a function approximation, $\tilde{f}$, since,

$$f(x) = \tilde{f}(x) + \varepsilon(x). \quad (42)$$

Although it is possible to obtain highly accurate estimates for the errors in particular models (see e.g. Oberkampf & Roy 2010), it is crucial to note that any form of practical on-the-fly uncertainty

[2]Note, however, that the existence of the inverses of $\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}$ and $\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\Phi}$ still depends on the choice of basis functions and the positions of the data points.

quantification will always only be an approximation for the true error. Just as the quality of the approximation highly depends on the estimation method, so does the quality of the error.

#### *2.3.1 Uncertainty in the probabilistic numerical paradigm*

Following the probabilistic numerical paradigm (Hennig et al. 2015; Cockayne et al. 2019; Hennig et al. 2022), we aim to quantify the uncertainty in the solution of linear operator equations by modelling the distribution over possible solutions conditioned on the data. In particular, we will use the expectation of the conditioned distribution as our function approximation,

$$\tilde{f}(x) \equiv \mathbb{E}[F(x) \mid \boldsymbol{y}]. \tag{43}$$

As a result, we can estimate the expected squared error in this approximation with the variance of the conditioned distribution,

$$\tilde{\varepsilon}^2(x) \equiv \mathbb{V}[F(x) \mid \boldsymbol{y}]. \tag{44}$$

This can be inferred from the fact that the stochastic function, $F(x)$, with corresponding stochastic error, $\mathcal{E}(x)$, ought to be related as

$$F(x) = \tilde{f}(x) + \mathcal{E}(x), \tag{45}$$

and the definition of the variance, which implies that

$$\mathbb{V}[F(x) \mid \boldsymbol{y}] = \mathbb{E}\left[\left(F(x) - \tilde{f}(x)\right)^2 \mid \boldsymbol{y}\right] = \mathbb{E}\left[\mathcal{E}(x)^2 \mid \boldsymbol{y}\right]. \tag{46}$$

Assuming that the probabilistic model, $F(x) \mid \boldsymbol{y}$, is an adequate model for the actual function, $f(x)$, the variance thus quantifies the expected squared error in the function approximation. Note that in our particular case, where the posterior is a Gaussian, the variance does not depend on the function values, $\boldsymbol{y}$, of the data but only on the locations at which the function was evaluated, $\boldsymbol{x}$.

Based on the variance in the dual formulation (32), one can derive an upper and lower bound on the expected squared error,

$$0 \leq \tilde{\varepsilon}^2(x) \leq k(x, x), \tag{47}$$

where, in the left inequality, we used that the variance has to be positive and in the right inequality that $k(\boldsymbol{x}, \boldsymbol{x}) + \boldsymbol{\sigma}^2$ is a positive definite matrix, such that the second term in (32) is always negative. It might seem odd to have an error measure that is bounded from above. However, one should note that it is not an upper bound on the actual error, but rather an upper bound on the expected error.

There is an alternative way to understand the error measure defined in (44) using the reproducing kernel Hilbert space (RKHS) of the kernel (see e.g. Cockayne et al. 2017). Let $\mathcal{H}$ denote the RKHS of the kernel defined in (7), with an associated inner product, $\langle\rangle_{\mathcal{H}}$, and norm $\|\cdot\|_{\mathcal{H}} \equiv \sqrt{\langle\cdot,\cdot\rangle_{\mathcal{H}}}$. If we now consider the projection $Pf \in \mathcal{H}$ of the function, $f$, in the RKHS, $\mathcal{H}$, one can derive the following bound (see Appendix A5),

$$\left|Pf(x) - \tilde{f}(x)\right| \leq \|Pf\|_{\mathcal{H}}\, \tilde{\varepsilon}(x). \tag{48}$$

This means that $\tilde{\varepsilon}(x)$ bounds the local error in the approximation, $\tilde{f}$, as measured in the RKHS. To intuitively see how this comes about without needing the notion of a RKHS, note that if the function that one tries to approximate is $f(x) = k(\boldsymbol{x}, x)$, based on the data $\mathbf{y} = k(\boldsymbol{x}, \boldsymbol{x})$, then the variance (32) is exactly equal to the difference between $f$, and its approximation (31). Now if the function one tries to approximate is not $k(\boldsymbol{x}, x)$, but a linear combination of evaluations of $k(\boldsymbol{x}, x)$, this difference can grow by an additional factor which can be bounded by $\|Pf\|_{\mathcal{H}}$, yielding the bound (48).

It should be emphasized that inequality (48) only bounds the error in the approximation with respect to the projection of the true solution in the RKHS, $Pf$, and not the error with respect to the true solution, $f$, itself. Hence, the strength of this bound crucially depends on the RKHS, and thus on the particular kernel, or equivalently, on the particular set of basis functions that is used. If the projection, $Pf$, in the RKHS is a good approximation for the true function, $f$, then the error bound (48) can also be used to bound the true error in equation (42). However, this assumes a certain regularity of the function, $f$, and again crucially depends on the particular choice of kernel or basis functions.

If $\{\phi_i\}$ is a finite and orthonormal set of square-integrable basis functions on some domain, $D$, i.e. $\langle\phi_i, \phi_j\rangle \equiv \int_D \phi_i \phi_j = \delta_{ij}$, then the function space spanned by these basis functions is a RKHS, say $\mathcal{H}$, with reproducing kernel (7), with respect to the following inner product. Since every function in the RKHS can be expressed as a linear combination of the basis functions, the inner product between $g(x) = \boldsymbol{g}^{\mathrm{T}}\boldsymbol{\phi}(x) \in \mathcal{H}$ and $h(x) = \boldsymbol{h}^{\mathrm{T}}\boldsymbol{\phi}(x) \in \mathcal{H}$ can be defined as $\langle g, h\rangle_{\mathcal{H}} \equiv \boldsymbol{g}^{\mathrm{T}}\boldsymbol{\lambda}^{-2}\boldsymbol{h}$. Hence, for a finite set of orthonormal basis functions, it is this inner product that must be used to compute the norm in the local error bound (48).

### 2.4 Example basis functions and kernels

Given the data, the key parameters that determine the regression model are the set of basis functions or the kernel in the primal and dual formulation respectively. In order to gain more insight into the linear regression method, we consider some specific examples of basis functions and their corresponding kernels.

#### *2.4.1 Fourier basis*

As a first example, consider the set of $N_{\mathrm{b}} = 2N + 1$ real Fourier basis functions, $\{1\} \cup \{\sin(\omega_n x)\}_{n=1}^{N} \cup \{\cos(\omega_n x)\}_{n=1}^{N}$, where we defined $\omega_n \equiv 2\pi n/L$, with $L$ the size of the domain that we are interested in. Given these basis functions, the primal representation of the function approximation (5) corresponds to the (truncated) Fourier series of the function that we are looking for. If we denote the entry in $\boldsymbol{\lambda}$ corresponding to the constant with $\lambda$, the entries corresponding to the sines with $\lambda_n$, and the entries corresponding to the cosines with $\lambda_n'$, the resulting kernel can be expanded as

$$\begin{aligned} k(x, x') = \lambda^2 &+ \sum_{n=1}^{N} \left(\frac{\lambda_n^2 + \lambda_n'^2}{2}\right) \cos\left(\omega_n(x - x')\right) \\ &+ \sum_{n=1}^{N} \left(\frac{\lambda_n^2 - \lambda_n'^2}{2}\right) \cos\left(\omega_n(x + x')\right). \end{aligned} \tag{49}$$

Typically, one would expect that modes corresponding to the same length scale, i.e. same $n$, would have similar weights, $\lambda_n \approx \lambda_n'$, such that the second summation vanishes. This approximately renders the kernel into a radial basis function $k(x, x') \approx K(\|x - x'\|)$.

Now consider the simplest case, when all $\lambda_n = \lambda_n' = \lambda$. The kernel is then a radial basis function and can be computed explicitly,

$$k(x, x') = \frac{\lambda^2}{2}\left(1 + \frac{\sin\left(\pi(2N+1)(x-x')/L\right)}{\sin\left(\pi(x-x')/L\right)}\right), \tag{50}$$

which is commonly known as (one half plus) the Dirichlet kernel. The kernel attains its maximum on the diagonal, $k(x, x) = \lambda^2(N + 1)$, and oscillates and decays away from there. The dual solution (11) is a linear combination of these radial basis functions centred, and thus peaking, around the data points, and decaying away elsewhere.

Note that, as the number of basis functions increases, $N \to \infty$, the kernel becomes more narrow and peaked, and in the limit tends

towards a delta distribution. Since the kernel centred around a data point represents the influence of that data point on the solution, this implies that with increasing $N$, the effect of each individual data point on the solution decreases and becomes ever more confined to a shrinking region around each data point. Similarly, with increasing $N$, the variance (or corresponding uncertainty estimate) in between data points will increase. Intuitively, this can be understood, since increasing the number of basis functions, while keeping the number of data points fixed, will imply that the basis functions are ever less constrained by the data, a problem commonly known as over-fitting.

Over-fitting can be cured with regularization by damping the higher order modes in the kernel through $\boldsymbol{\lambda}$, making it less peaked in the limit of large $N$. Note that the entries of the regularization vector, $\boldsymbol{\lambda}$, appear as the Fourier coefficients of the kernel (49). This illustrates the crucial interplay between the choice of basis functions and regularization. In a sense, regularization effectively comes down to re-scaling the basis functions, since by making the re-scaling, $\phi_i \rightarrow \lambda_i \phi_i$, the regularization vector can always be cast into the trivial form $\boldsymbol{\lambda} = \mathbb{1}$.

The Fourier basis allows us to restrict the desired solution of the regression problem to a minimal length scale, defined by $L/N$. Therefore, by considering only a limited number of basis functions, $N_\mathrm{b} \ll N_\mathrm{d}$, one can obtain a large-scale ($L/N$) approximation for the model, which can efficiently be solved in the primal formulation, as it only requires the solution of an $(N_\mathrm{b} \times N_\mathrm{b})$-dimensional system.

A potential problem with the Fourier basis is that the effective width, $\xi$, of the kernel (49), which can be estimated as, $\xi \sim L/N$, is the same around every data point. This is fine as long as the distance between the data points, $\boldsymbol{x}$, is much smaller than the effective width of the kernel, but causes problems, for instance, if there are 'lonely' data points whose nearest neighbours are much farther away than the effective width of the kernel. Since the kernel rapidly decays for length scales beyond its effective width, the solution around a so-called 'lonely' data point, say $x_l$, can be approximated as,

$$\tilde{f}(x) \;=\; \boldsymbol{v}^\mathrm{T} k(\boldsymbol{x}, x) \;\approx\; v_l k(x_l, x), \tag{51}$$

which holds, as long as $x$ is much closer to $x_l$ than to its nearest neighbour, say $x_l^\mathrm{n}$, i.e. $\|x - x_l\| \ll \|x - x_l^\mathrm{n}\|$. Since the nearest neighbour is much farther away than the effective width of the kernel (by definition of a 'lonely point'), there is a significant region around $x_l$, defined by $\{x : \xi < \|x - x_l\| \ll \|x - x_l^\mathrm{n}\|\}$, for which,

$$\tilde{f}(x) \;\approx\; 0. \tag{52}$$

Similarly, the variance (or uncertainty estimate) for the function approximation in that region will attain its maximum value,

$$\tilde{\varepsilon}^2(x) \;\approx\; k(x, x) \;=\; \lambda^2 (N + 1). \tag{53}$$

Hence, the approximation is probably not good in that region. This type of problem can be avoided by choosing basis functions or a kernel that is locally adapted to the distribution of data points.

### 2.4.2 Radial basis functions

As a second example, consider a set of basis functions generated by a radial basis function, $\psi$, centred around each data point, $x_i$, such that there is one basis function, $\phi_i$ for each data point $x_i$, with,

$$\phi_i(x) \;=\; \psi\left(\frac{\|x - x_i\|}{\xi_i}\right), \tag{54}$$

in which $\xi_i$ controls the effective width of the radial basis function around data point, $x_i$. The corresponding kernel for this basis reads

$$k(x, x') \;=\; \sum_{i=1}^{N_\mathrm{b}} \psi\left(\frac{\|x - x_i\|}{\xi_i}\right) \lambda_i^2 \, \psi\left(\frac{\|x' - x_i\|}{\xi_i}\right). \tag{55}$$

Issues with 'lonely' data points as encountered with the Fourier basis can be avoided here by tailoring the basis functions to the data, for instance, by choosing $\xi_i = \|x_i - x_i^\mathrm{n}\|$, in which $x_i^\mathrm{n}$ is the nearest neighbour of $x_i$.

Radial basis functions are a popular choice for solving operator equations. Although, often, some care is required to cope with the ill-conditioning of the resulting linear system (see e.g. Fornberg & Flyer 2015). Moreover, radial basis functions often offer an intuitive interpretation. For instance, when dealing with smoothed-particle hydrodynamics data (Gingold & Monaghan 1977; Lucy 1977), the basis functions can be related to the smoothing kernels and represent the proliferation of the data from each particle.

Since, in this approach, the basis functions are tied to the data points, approximating the solution around certain data points can be achieved by discarding the corresponding basis functions. In De Ceuster et al. (2020), a mesh reduction method for radiative transfer models was proposed in which, based on a certain heuristic, data points which were not deemed essential were discarded from the model. A similar but improved reduction scheme can be obtained using the linear regression approach with basis functions tied to the data, by discarding the corresponding basis functions instead. In this way, the data itself does not have to be discarded and can still be taken into account, while the model is reduced in computational complexity. Furthermore, the Bayesian linear regression method can provide an estimate for the uncertainty on the result after solving the reduced model. This leads us to the question whether there are even better bases to compress radiative transfer models.

### 2.4.3 Wavelet bases

Wavelets have a proven track record for data compression in various applications, such as sound and image processing (see e.g. Vetterli 2001), and have already successfully been applied to solve operator equations (see e.g. Stevenson 2009, and the references therein). They combine the localization in scale of the Fourier basis, i.e. certain basis functions describe certain length scales, with the localization in space of radial basis functions, i.e. certain basis functions describe certain regions in space. As a result, wavelet bases are of the form,

$$\psi_{mn}(x) \;=\; a^{-m/2} \, \psi\left(\frac{x - nb}{a^m}\right), \tag{56}$$

indexed by two indices, in which $m$ describes the length scale, and $n$ describes the location, parametrized by the constants $a$ and $b$ respectively. By imposing the mathematical structure of a multiresolution analysis, relations between the different scales can be derived which allow one to construct orthogonal wavelet bases, which give rise to efficient algorithms to decompose functions into their wavelet components (see e.g. Daubechies 1992).

By selecting (or disregarding) certain wavelet basis functions we thus can locally refine (or coarsen) the solution of the linear regression problem. However, this assumes that we know where we want to refine the model and where a coarser representation suffices. Alternatively, by expressing the data directly in a wavelet basis (e.g. using a fast wavelet transform), one can select only those components that significantly contribute (see e.g. Daubechies 1992).

There is a large variety of wavelet bases and the key remains to choose an appropriate one. Moreover, when the aim is to solve

operator equations, the wavelet basis should still be appropriate when acted upon with the relevant operator. Choosing appropriate wavelet bases adapted to a particular operator turns out to be a challenging endeavour (see e.g. Stevenson 2009). However, recently, significant progress has been made, for instance, by Owhadi (2017), which makes wavelet bases an attractive choice for solving large (Bayesian) linear regression problems (see also Section 4).

## 3 BAYESIAN RADIATIVE TRANSFER

We can now apply the probabilistic numerical approach developed above to the particular case of radiative transfer problems. The goal is to find the radiation field throughout a region, based on the radiative properties of the medium and some boundary conditions. The radiation field can be described by the specific monochromatic intensity, $I_\nu(x, \hat{n})$, i.e. the energy at a point, $\boldsymbol{x}$, transported in a direction, $\hat{n}$, in a certain frequency bin, $\nu$. Interactions between the radiation field and the medium can be described in terms of the change they imply in the specific monochromatic intensity. The radiative transfer equation is a linear operator equation that relates this change to the radiative properties of the medium,

$$\mathscr{L} I_\nu(x, \hat{n}) \;=\; \eta_\nu(x), \tag{57}$$

in which, $\eta_\nu(\boldsymbol{x})$, is the emissivity of the medium. In the time-independent case and including scattering, the operator, $\mathscr{L}$, acts on the intensity as (see e.g. Mihalas & Weibel-Mihalas 1984),

$$\mathscr{L} I_\nu(x, \hat{n}) \equiv \left(\chi_\nu(x) \;+\; \hat{n}\cdot\nabla\right) I_\nu(x, \hat{n}) - \oint \mathrm{d}\Omega' \int_0^\infty \mathrm{d}\nu'\; \Phi_{\nu\nu'}\left(x, \hat{n}, \hat{n}'\right) I_{\nu'}(x, \hat{n}'). \tag{58}$$

Here, we introduced the opacity, $\chi_\nu(\boldsymbol{x})$, and the scattering redistribution function, $\Phi_{\nu\nu'}\left(x, \hat{n}, \hat{n}'\right)$. Since $\mathscr{L}$ is a linear operator, the solution of the radiative transfer equation, given appropriate boundary conditions, can be viewed as a Bayesian linear regression problem. It remains to find an appropriate set of basis functions (in the primal formulation), or to find an appropriate kernel (in the dual formulation), given the radiative properties of the medium.

### 3.1 Approximate radiative transfer models

Almost all astrophysical simulations require some kind of radiative transfer model. However, due to the significant computational cost, one is often forced to make drastic approximations. In this section, we show how the primal formulation can be used to create reduced-order or approximate radiative transfer models and show how it can be applied, for instance, to compute approximated Lambda operators for atomic and molecular line transfer.

#### *3.1.1 Reduced-order models*

As already alluded to in Section 2.4, we can obtain approximate solutions for a linear regression problem by considering reduced sets of basis functions in the primal formulation. The basis functions essentially map the regression problem to an $N_\mathrm{b}$-dimensional feature space in which the problem is solved. Therefore, in a sense, the primal solution (5) can be interpreted as follows:

$$\tilde{f}(x) \;=\; \boldsymbol{y}^\mathrm{T} \underbrace{\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}}_{\text{compress}} \underbrace{\left(\boldsymbol{\Phi}^\mathrm{T}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi} + \boldsymbol{\lambda}^{-2}\right)^{-1}}_{\text{solve}} \underbrace{\boldsymbol{\phi}(x)}_{\text{decompress}}. \tag{59}$$

First, the $N_\mathrm{d}$-dimensional data vector, $\boldsymbol{y}$, is mapped into the $N_\mathrm{b}$-dimensional feature space, which can be viewed as a projection or compression, if $N_\mathrm{b} < N_\mathrm{d}$. Then, the problem is solved in the $N_\mathrm{b}$-dimensional feature space, and finally mapped back into the desired format. The least-squares problem posed in equation (3) minimizes the compression loss. The resulting reduced-order model provides an approximate solution to the (more) exact radiative transfer problem, in contrast to the exact solutions to approximate models that are often used. Moreover, the probabilistic interpretation allows us to quantify with the variance (28) the uncertainty that was introduced by compressing the model, allowing us to strictly control the trade-off between accuracy and computational cost.

By denoting the first part of equation (59) as a compression of the data, one could ask whether the vector-matrix multiplication in equation (59) is the most efficient way to perform this compression. Indeed, for the Fourier and wavelet bases there exist more efficient algorithms to express a given data set into these bases, the so-called Fast Fourier Transform (FFT) and Fast Wavelet Transform (FWT), respectively (see e.g. Press et al. 2007). These can reduce the computational cost of these models even further.

The type and amount of compression critically depends on the set of basis functions that is used. One way to choose them, for instance, is by performing a principle component analysis on the design matrix, $\boldsymbol{\Phi}$. This yields what is known as a proper orthogonal decomposition (POD; see e.g. Benner et al. 2017). In addition to performing this compression, the probabilistic approach now also allows to quantify the uncertainties that are thus introduced.

#### *3.1.2 Application: approximate Lambda operators*

Approximations to radiative transfer are often used to accelerate iterative line radiative transfer solvers. Models involving atomic or molecular line radiative transfer show a non-linear coupling between the radiative properties of the medium and the radiation field. This coupling can be expressed as

$$I \;=\; \Lambda\left[\eta\left(I\right)\right], \tag{60}$$

in which $I$ indicates the radiation field, $\Lambda$ is a linear operator, and we explicitly indicated the dependence of $\eta$ on the radiation field. It is this dependency of $\eta$ on $I$ that is usually non-linear. Due to this non-linear coupling, the radiation field has to be computed in an iterative way, (see e.g. chapter 13 in Hubeny & Mihalas 2014),

$$I^{(n+1)} \;=\; \Lambda\left[\eta\left(I^{(n)}\right)\right]. \tag{61}$$

This iterative scheme often shows notoriously slow convergence and one often has to resort to acceleration techniques, such as operator splitting (Cannon 1973a,b), which yields the implicit scheme,

$$I^{(n+1)} \;=\; \Lambda^*\left[\eta\left(I^{(n+1)}\right)\right] \;+\; \left(\Lambda - \Lambda^*\right)\left[\eta\left(I^{(n)}\right)\right], \tag{62}$$

in which the linear operator, $\Lambda^*$, is an approximation for the operator, $\Lambda$, that can easily be inverted (see e.g. Rybicki & Hummer 1991, for a specific implementation). Intuitively, the better the approximation, $\Lambda^*$, the smaller the dependence on the previous iteration in (62), and thus the better convergence will be. The key to success in this acceleration scheme is to find a good approximate operator, $\Lambda^*$.

Comparing equations (5), (57), and (60), one can see that, in the primal formulation, the operator, $\Lambda$, in matrix form is given by

$$\boldsymbol{\Lambda} \;=\; \boldsymbol{\phi}(x)^\mathrm{T}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T}\left(\boldsymbol{\Phi}\boldsymbol{\lambda}^2\boldsymbol{\Phi}^\mathrm{T} + \boldsymbol{\sigma}^2\right)^{-1}. \tag{63}$$

As a result, good approximations to this operator can be obtained, for instance, by considering reduced sets of basis functions in the corresponding linear regression problem, as shown in Section 3.1.1.

### 3.2 Method of characteristics

In order to make the probabilistic approach to radiative transfer more concrete, we consider the specific example of the method of characteristics and derive it from a Bayesian point of view.

In its simplest form, in the absence of scattering and neglecting any frequency dependence, the time-independent radiative transfer equation along a single ray reads

$$\mathscr{L}_s I(s) \;=\; \eta(s), \tag{64}$$

in which the linear differential operator, $\mathscr{L}_s$, is defined as

$$\mathscr{L}_s \;\equiv\; \chi(s) \;+\; \partial_s. \tag{65}$$

For future reference, we already note that the Green's function for this linear operator, $\mathscr{L}_s$, is given by

$$G(z,\,s) \;=\; \Theta(s-z)\, e^{-\tau(z,\,s)}, \tag{66}$$

in which $\Theta$ is the Heaviside function, and the optical depth, $\tau$, over an interval $[z,\,s]$ along the ray, is defined as

$$\tau(z,\,s) \;\equiv\; \int_z^s \mathrm{d}s'\, \chi(s'), \tag{67}$$

such that $\partial_s \tau(z,\,s) = \chi(s)$, and thus, as expected,

$$\mathscr{L}_s G(z,\,s) \;=\; \delta(s-z). \tag{68}$$

Using this Green's function one can (at least formally) solve the radiative transfer equation, as in the method of characteristics.

#### 3.2.1 *Classical method of characteristics*

The method of characteristics solves the transfer equation starting from its formal solution based on the Green's function. Given the boundary condition, $I(s_0) = I_0$, at boundary point, $s_0$, one finds

$$I(s) \;=\; I_0\, e^{-\tau(s_0,s)} \;+\; \int_{s_0}^{s} \mathrm{d}s'\, \eta(s')\, e^{-\tau(s',s)}. \tag{69}$$

The required integrals in equations (67) and (69) are then evaluated using a (local) interpolation both for the emissivity and opacity functions, $\eta(s)$ and $\chi(s)$.

At this point, a distinction is often made between so-called short and long characteristic methods depending on the location of the point $s_0$ in the discretization. In the case of short characteristics, $s_0$ is taken to be the previous point in the discretization, while, in the case of long characteristics, it is taken to be the boundary of the computational domain. For our intents and purposes this distinction does not matter, so we continue with the formulation in (69), in which $s_0$ can be any point in the discretization.

The emissivity and opacity are usually interpolated using a linear scheme. Given a kernel, $\kappa$, the interpolant (11) in the dual formulation can be written as

$$\tilde{\eta}(s) \;\equiv\; \boldsymbol{\eta}^{\mathrm{T}} \left(\kappa(\boldsymbol{a},\boldsymbol{a}) + \boldsymbol{\sigma}_\eta^2\right)^{-1} \kappa(\boldsymbol{a},s), \tag{70}$$

$$\tilde{\chi}(s) \;\equiv\; \boldsymbol{\chi}^{\mathrm{T}} \left(\kappa(\boldsymbol{a},\boldsymbol{a}) + \boldsymbol{\sigma}_\chi^2\right)^{-1} \kappa(\boldsymbol{a},s), \tag{71}$$

in which $\boldsymbol{a}$ is the vector of positions at which the values for $\eta$ and $\chi$ are given, and where $\boldsymbol{\sigma}_\eta^2$ and $\boldsymbol{\sigma}_\chi^2$ denote the diagonal matrices with the variances for the given values of $\eta$ and $\chi$ respectively. However, in the classical method of characteristics, these variances are never used, and thus implicitly assumed to be negligible. In the Bayesian method of characteristics, however, they model the uncertainties in the emissivities and opacities that originate, for instance, from the uncertainties in the radiative data (see Section 3.2.2). Furthermore, we note that, in principle, one can use a different kernel for $\eta$ and $\chi$, although, in practice, one often uses the same one. One particularly popular choice of kernel is the one corresponding to the basis of Lagrange polynomials, since they trivially satisfy the interpolation property. With equations (70) and (71), the formal solution yields

$$\tilde{I}(s) \;=\; I_0\, e^{-\tilde{\tau}(s_0,s)} \;+\; \boldsymbol{\eta}^{\mathrm{T}}\, \mathbf{K}_\eta^{-1} \int_{s_0}^{s} \mathrm{d}s'\, \kappa(\boldsymbol{a},s')\, e^{-\tilde{\tau}(s',s)} \tag{72}$$

in which the interpolated optical depth is given by

$$\tilde{\tau}(z,\,s) \;=\; \boldsymbol{\chi}^{\mathrm{T}} \mathbf{K}_\chi^{-1} \int_z^s \mathrm{d}s'\, \kappa(\boldsymbol{a},s'), \tag{73}$$

where, for brevity, we defined the matrices $\mathbf{K}_\eta \equiv \kappa(\boldsymbol{a},\boldsymbol{a}) + \boldsymbol{\sigma}_\eta^2$, and $\mathbf{K}_\chi \equiv \kappa(\boldsymbol{a},\boldsymbol{a}) + \boldsymbol{\sigma}_\chi^2$. The integrals in equations (72) and (73) can now be evaluated on the (analytically) known kernel function, $\kappa$, thus solving the radiative transfer equation.

#### 3.2.2 *Bayesian method of characteristics*

Now we show how the method of characteristics can be derived as a Bayesian linear regression problem in the dual formulation by choosing a particular type of kernel, or equivalently by choosing a particular set of basis functions.

Given the Green's function (66) for the differential operator in the radiative transfer equation, consider a kernel of the form,

$$k(z,\,s) \;=\; \int_{-\infty}^{+\infty} \mathrm{d}s' \int_{-\infty}^{+\infty} \mathrm{d}z'\, \kappa(s',\,z')\, G(z',\,z)\, G(s',\,s), \tag{74}$$

in which $\kappa(s',\,z')$ is another kernel from which we only demand that it does not correlate the region $s > s_0$ with $s < s_0$. The reason for this is, that, in the classical method of characteristics, we want to use the solution at $s_0$ as a true boundary condition, i.e. such that nothing at $s > s_0$ affects the solution at $s < s_0$, and vice versa. This implies a block diagonal kernel of the form,

$$\begin{aligned}\kappa(z,\,s) \equiv\; & \Theta(s_0-z)\Theta(s_0-s)\kappa(z,\,s) \\ & + \Theta(z-s_0)\Theta(s-s_0)\kappa(z,\,s).\end{aligned} \tag{75}$$

If we now assume that $\forall a \in \boldsymbol{a} : a > s_0$, and we assume no error on the boundary condition, one can show (see Appendix A6) that the dual solution for the Bayesian linear regression problem reads

$$\tilde{I}(s) \;=\; I_0\, e^{-\tilde{\tau}(s_0,s)} \;+\; \boldsymbol{\eta}^{\mathrm{T}}\, \mathbf{K}_\eta^{-1} \int_{s_0}^{s} \mathrm{d}s'\, \kappa(\boldsymbol{a},s')\, e^{-\tilde{\tau}(s',s)} \tag{76}$$

with the corresponding uncertainty estimate given by

$$\begin{aligned}\tilde{\varepsilon}_{\tilde{I}}^2(s) = & \int_{s_0}^{s} \mathrm{d}s' \int_{s_0}^{s} \mathrm{d}z'\, e^{-\tilde{\tau}(s',\,s)} e^{-\tilde{\tau}(z',\,s)} \\ & \times \left(\kappa(s',\,z') - \kappa(\boldsymbol{a},\,s')^{\mathrm{T}} \mathbf{K}_\eta^{-1} \kappa(\boldsymbol{a},\,z')\right).\end{aligned} \tag{77}$$

Note that the probabilistic solution (76) is exactly the same as the classical solution (72) for the method of characteristics. Therefore, we can conclude that both methods are equivalent, but with the important difference that the probabilistic approach can account for uncertainties on the input (through $\boldsymbol{\sigma}_\eta$) and we thus can estimate the uncertainty on the result. Moreover, in the expression between parentheses in equation (77),

$$\kappa(s',\,z') - \kappa(\boldsymbol{a},\,s')^{\mathrm{T}} \mathbf{K}_\eta^{-1} \kappa(\boldsymbol{a},\,z') \tag{78}$$

we recognize the resulting variance in the dual formulation (32) that stems from the interpolation of the emissivity (70).

We should note that in the definition of the kernel (74), we implicitly assumed that we knew the Green's function (66), and thus we implicitly assumed that we knew the optical depth (67). In

general, we do not have an exact expression for the optical depth. However, we can find an approximate solution by solving another Bayesian linear regression problem for the operator equation,

$$\partial_s \tau(z,\, s) \;=\; \chi(s), \tag{79}$$

with boundary condition, $\tau(z,\, z) = 0$, which, using the kernel,

$$k(z,\, s) \;=\; \int_{-\infty}^{+\infty} \mathrm{d}s' \int_{-\infty}^{+\infty} \mathrm{d}z'\, \kappa(s', z'), \tag{80}$$

unsurprisingly, yields the expected solution,

$$\tilde{\tau}(z,\, s) \;=\; \boldsymbol{\chi}^{\mathrm{T}} \mathbf{K}_{\chi}^{-1} \int_z^s \mathrm{d}s'\, \kappa(\boldsymbol{a}, s'), \tag{81}$$

with the corresponding uncertainty estimate given by

$$\tilde{\varepsilon}_\tau^2(z,\, s) = \int_z^s \mathrm{d}s' \int_z^s \mathrm{d}z' \times \left(\kappa(s',\, z') - \kappa(\boldsymbol{a},\, s')^{\mathrm{T}} \mathbf{K}_{\chi}^{-1} \kappa(\boldsymbol{a},\, z')\right). \tag{82}$$

This can now be used to define the Green's function (66).

It should be emphasized that the uncertainty on the optical depth, and by extension the uncertainty on the opacity, is not yet included in the uncertainty estimate for the radiation field (77). The expression (77) only includes the uncertainties on the emissivity, and not on the opacity or optical depth, because the opacity (only) appears in the linear operator (65), which in the Bayesian linear regression method is assumed to be deterministic. The reason for this is that imposing a probability distribution also on the linear operator would render the posterior distribution non-Gaussian, which would severely complicate conditioning and impede analytical solutions.

Nevertheless, in this particular case, an analytic solution can still be obtained for the expectation and the variance of the radiation field, taking into account the distribution of the opacity, although the resulting distribution is not a Gaussian anymore.

From equations (76) and (77), and equations (81) and (82), we know the distributions of the stochastic functions $I$ and $\tau$,

$$p\left(I \,|\, \tau\right) \;=\; \mathcal{N}\left(\tilde{I},\, \tilde{\varepsilon}_I^2\right), \tag{83}$$

$$p\left(\tau\right) \;=\; \mathcal{N}\left(\tilde{\tau},\, \tilde{\varepsilon}_\tau^2\right). \tag{84}$$

The expectation, $\hat{I}$, and variance, $\hat{\varepsilon}_I^2$, of the radiation field with respect to the joint distribution with $\tau$ can then be obtained using the law of total expectation (see Appendix A7),

$$\hat{I} \;\equiv\; \mathbb{E}\left[I\right] \;=\; \mathbb{E}_\tau\left[\mathbb{E}\left[I \,|\, \tau\right]\right] \;=\; \mathbb{E}_\tau\left[\tilde{I}\right], \tag{85}$$

and similarly, using the law of total variance (see Appendix A8),

$$\begin{aligned}\hat{\varepsilon}_I^2 \;&\equiv\; \mathbb{V}\left[I\right] = \mathbb{E}_\tau\left[\mathbb{V}\left[I \,|\, \tau\right]\right] \;+\; \mathbb{V}_\tau\left[\mathbb{E}\left[I \,|\, \tau\right]\right] \\ &=\; \mathbb{E}_\tau\left[\tilde{\varepsilon}_I^2\right] \;+\; \mathbb{E}_\tau\left[\tilde{I}^2\right] \;-\; \hat{I}^2.\end{aligned} \tag{86}$$

Evaluating these expectations yields (see Appendix A9),

$$\hat{I}(s) \;=\; I_0\, e^{-\hat{\tau}(s_0,s)} \;+\; \boldsymbol{\eta}^{\mathrm{T}} \mathbf{K}_{\eta}^{-1} \int_{s_0}^{s} \mathrm{d}s'\, \kappa(\boldsymbol{a}, s')\, e^{-\hat{\tau}(s',s)} \tag{87}$$

with the corresponding uncertainty estimate given by

$$\begin{aligned}\hat{\varepsilon}_I^2(s) \;\leq\; &\int_{s_0}^{s} \mathrm{d}s' \int_{s_0}^{s} \mathrm{d}z'\, e^{-\overline{\tau}(s',\,s)} e^{-\overline{\tau}(z',\,s)} \\ &\times \left(\kappa(s',\, z') - \kappa(\boldsymbol{a},\, s')^{\mathrm{T}} \mathbf{K}_{\eta}^{-1} \kappa(\boldsymbol{a},\, z')\right) \\ &+\; \bar{I}^2(s) \;-\; \hat{I}^2(s).\end{aligned} \tag{88}$$

These expressions look very similar to equations (76) and (77). The only difference is that the optical depth is replaced by newly defined effective optical depths,

$$\hat{\tau}(z,\, s) \;\equiv\; \tilde{\tau}(z,\, s) \;-\; \tfrac{1}{2}\, \tilde{\varepsilon}_\tau^2(z,\, s), \tag{89}$$

$$\overline{\tau}(z,\, s) \;\equiv\; \hat{\tau}(z,\, s) \;-\; \tfrac{1}{2}\, \tilde{\varepsilon}_\tau^2(z,\, s), \tag{90}$$

and there are additional terms in (87), which account for correlations in the optical depth. The intensity, $\bar{I}$, is defined analogously to $\hat{I}$, but with $\hat{\tau}$ replaced by $\overline{\tau}$. Equation (88) only gives a practical upper bound. In Appendix A9, we also derive the complete expression for $\hat{\varepsilon}_I^2$. The uncertainty on the optical depth thus causes an effective reduction of the optical depth that appears in the radiation field (87).

### 3.3 Example

We illustrate the Bayesian method of characteristics with a simple example. Consider a set of $N_{\mathrm{d}}$ points, $\{s_d\}$, at which we know the emissivities, $\{\eta_d\}$, and opacities, $\{\chi_d\}$. Moreover, consider a set of $N_{\mathrm{b}} = N_{\mathrm{d}}$ basis functions that satisfy the interpolation property for the data points, i.e. $\phi_i(s_d) = \delta_{di}$, such that the design matrix is an identity matrix. As a result, we have that $\kappa(\boldsymbol{a}, \boldsymbol{a}) = \boldsymbol{\lambda}^2$, such that, if we use the same interpolation scheme both for $\eta$ and $\chi$, we have that, $\mathbf{K}_\eta \equiv \boldsymbol{\lambda}^2 + \boldsymbol{\sigma}_\eta^2$, and $\mathbf{K}_\chi \equiv \boldsymbol{\lambda}^2 + \boldsymbol{\sigma}_\chi^2$. Furthermore, one can show that $\kappa(\boldsymbol{a}, s) = \boldsymbol{\lambda}^2 \boldsymbol{\phi}(s)$. Finally, we assume that $\boldsymbol{\lambda} = \lambda \mathbb{1}$, and we assume the same uncertainty for every data point, such that $\boldsymbol{\sigma}_\eta = \sigma_\eta \mathbb{1}$ and $\boldsymbol{\sigma}_\chi = \sigma_\chi \mathbb{1}$. The resulting optical depth and the corresponding uncertainty estimate can then be written as

$$\tilde{\tau}(z,\, s) \;=\; \tfrac{\lambda^2}{\lambda^2+\sigma_\chi^2}\, \boldsymbol{\chi}^{\mathrm{T}} \boldsymbol{\psi}_\tau(z, s), \tag{91}$$

$$\tilde{\varepsilon}_\tau^2(z,\, s) \;=\; \tfrac{\lambda^2\sigma_\chi^2}{\lambda^2+\sigma_\chi^2}\, \boldsymbol{\psi}_\tau(z, s)^{\mathrm{T}}\, \boldsymbol{\psi}_\tau(z, s), \tag{92}$$

in which $\boldsymbol{\chi}$ is the vector of opacities, and we defined the vector,

$$\boldsymbol{\psi}_\tau(z, s) \;\equiv\; \int_z^s \mathrm{d}s'\, \boldsymbol{\phi}(s'). \tag{93}$$

Similarly for the radiation field, we find that

$$\hat{I}(s) \;=\; I_0\, e^{-\hat{\tau}(s_0,s)} \;+\; \tfrac{\lambda^2}{\lambda^2+\sigma_\eta^2}\, \boldsymbol{\eta}^{\mathrm{T}} \hat{\boldsymbol{\psi}}_I(s), \tag{94}$$

$$\hat{\varepsilon}_I^2(s) \;\leq\; \tfrac{\lambda^2\sigma_\eta^2}{\lambda^2+\sigma_\eta^2}\, \bar{\boldsymbol{\psi}}_I(s)^{\mathrm{T}}\, \bar{\boldsymbol{\psi}}_I(s) \;+\; \bar{I}^2(s) \;-\; \hat{I}^2(s), \tag{95}$$

in which $\boldsymbol{\eta}$ is the vector of emissivities, and we defined the vectors,

$$\hat{\boldsymbol{\psi}}_I(s) \;\equiv\; \int_{s_0}^{s} \mathrm{d}s'\, \boldsymbol{\phi}(s')\, e^{-\hat{\tau}(s',\,s)}, \tag{96}$$

$$\bar{\boldsymbol{\psi}}_I(s) \;\equiv\; \int_{s_0}^{s} \mathrm{d}s'\, \boldsymbol{\phi}(s')\, e^{-\overline{\tau}(s',\,s)}. \tag{97}$$

Fig. 1 shows the solution the the radiative transfer equation along a single ray using the Bayesian method of characteristics. The variances were chosen comically large, especially to illustrate the effective optical depths. Even with such large variance, we see that the effect on the optical depth is relatively small. As basis functions, we choose the 25 fifth-order basis splines that interpolate the 25 uniformly distributed data points, i.e. such that $\phi_i(s_d) = \delta_{di}$. It should be emphasized that, although the probability distributions for the emissivity, opacity, and optical depth are all Gaussian, the probability distribution for the intensity is not a Gaussian. In fact, we have not specified the particular distribution, but could nevertheless determine the expectation and variance.

Although we only presented the probabilistic numerical method in the absence of scattering, neglecting any frequency dependence, and only along a single ray, we should note that it can readily by generalized to a three dimensions, including scattering and frequency dependence. How this can be done for a particular set of basis functions will be demonstrated in a forthcoming paper.

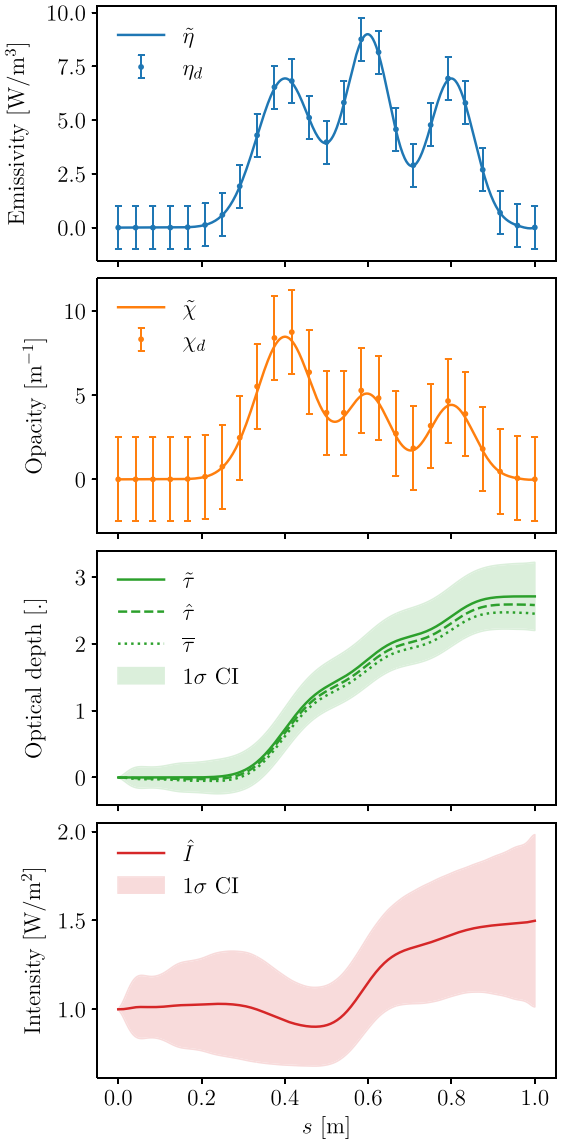


**Figure 1.** Example of the Bayesian method of characteristics for 25 data points, with $I_0 = 1.0$ W/m$^2$, $\sigma_\eta = 1.0$ W/m$^3$, $\sigma_\chi = 2.5$ m$^{-1}$, and $\lambda = 10.0$. The error bars on the data and the shaded areas around the curves show the respective (local) $1\sigma$ confidence intervals (CI), or its upper bound (95) for the intensity. The source code for this figure can be found at github.com/FredDeCeuster/RadiativeTransferAsRegression.

## 4 DISCUSSION

The probabilistic numerical method presented in this paper differs significantly from commonly-used (probabilistic) sampling-based Monte Carlo methods for uncertainty quantification (Metropolis & Ulam 1949). Where sampling-based Monte Carlo methods are non-intrusive and treat the physical model under consideration as a black box, the probabilistic numerical approach, as described here, requires to recast the entire description of the model as a Bayesian linear regression problem. This investment, however, pays off in a key advantage for the probabilistic numerical approach: where sampling-based methods typically require many model evaluations to obtain a distribution, the probabilistic numerical approach requires only a single, albeit computationally slightly more expensive, model evaluation. This is particularly advantageous for computationally expensive models, such as the ones encountered in radiative transfer, as described in this paper.

The (Bayesian) linear regression model is critically defined by the choice of basis functions or the choice of the corresponding kernel. As discussed in Section 2.4, different choices of basis functions give rise to different kinds of descriptions, or, when, $N_\mathrm{b} < N_\mathrm{d}$, different kinds of approximations. In particular, the truncation of a set of Fourier basis functions implies a characteristic minimal resolvable length scale, whether or not to include certain data-centred radial basis functions will alter the solution around those data points, and wavelets allow one to locally refine or coarsen the model. All of these particular bases have their particular advantages and disadvantages, but none of them is in any sense optimal. Furthermore, it should be noted that, in our discussion of different bases, we did not take into account the effect of the operator acting on the basis functions, while at the end of Section 2.1.4 it was emphasized that to solve a linear operator equation such as (13), the basis functions should ideally be chosen such that $\{\mathscr{L}\phi_i\}$ can properly fit $g(x)$, $\{\mathscr{B}\phi_i\}$ can properly fit $h(x)$, and $\{\phi_i\}$ can properly fit the sought after solution function $f(x)$. There are many different ways to solve this optimization problem of finding an appropriate (reduced) set of basis functions, often colloquially referred to as model-order reduction methods (see e.g. Benner et al. 2017). One particularly interesting method by Owhadi (2017) describes how to construct a basis that is in a sense optimal, based on probabilistic numerical considerations (see also Owhadi & Scovel 2019). In a forthcoming paper, we will choose a particular type of basis, show how it can be tailored to the problem at hand, and demonstrate how this can be used to solve radiative transfer problems in a practical three-dimensional setting.

Probabilistic numerical methods are by no means restricted to radiative transfer applications and can readily be applied to various other solvers of operator equations. In particular, we envision similar techniques to be useful, for instance, in chemical kinetics models that simulate the chemical evolution of a set of species, given a network of chemical reactions (see e.g. McElroy et al. 2013). These chemical networks are also often reduced to lower the computational cost (see e.g. Grassi et al. 2021). As a result, the probabilistic numerical setting might also there lead to interesting approximation techniques. However, since these are initial value problems, the required probabilistic numerical approach will probably be different from what was presented here and more along the lines of, for instance, Conrad et al. (2017).

## 5 CONCLUSION

Inspired by the probabilistic numerical approaches advocated, amongst others, by Hennig et al. (2015, 2022) and Cockayne et al. (2019), we have presented a way to view radiative transfer as a Bayesian linear regression problem. Specifically, we have modelled the solution of a radiative transfer problem with the expectation of a multivariate Gaussian probability distribution over possible solutions, conditioned on evaluations of the radiative transfer equation and boundary conditions. This allowed us to model uncertainties, both on the input and output of the model, with the variances of the associated probability distributions, without the need for computationally expensive (Monte Carlo) sampling schemes. Moreover, this method naturally allowed us to create reduced-order radiative transfer models, for which the probabilistic interpretation furthermore allowed us to quantify the uncertainty that was introduced by reducing the model. As an example, we showed how the commonly-used method of characteristics can be derived from a probabilistic point of view.

The aim of this paper was not to present the definitive probabilistic numerical approach for radiative transfer, but rather to motivate future research in this direction by showing the potential benefits of a probabilistic point of view and indicate connections with other research, for instance, by quantifying uncertainties from a statistical point or view, and viewing model reduction as a form of data compression.

## ACKNOWLEDGEMENTS

We would like to thank Maarten Baes and Peter Camps for planting the idea that the future of radiative transfer probably lies somewhere in between the typical probabilistic and deterministic approaches. We also thank Amery Gration and Johan Suykens for insightful discussions on Bayesian linear regression and Gaussian processes. Finally, we would also like to thank the scientific editor and the two anonymous reviewers for their considerate and helpful feedback. FDC is supported by the EPSRC iCASE studentship programme (ref. 1878976) and Intel Corporation. FDC and LD acknowledge support from the ERC consolidator grant 646758 AEROSOL. TC is a PhD fellow of the Research Foundation – Flanders (FWO).

## DATA AVAILABILITY

No new data were generated or analysed in support of this research.

## REFERENCES

Benner P., Ohlberger M., Cohen A., Willcox K., 2017, Model Reduction and Approximation. Society for Industrial and Applied Mathematics, Philadelphia, PA doi:10.1137/1.9781611974829, http://epubs.siam.org/doi/book/10.1137/1.9781611974829
Berlinet A., Thomas-Agnan C., 2004, Reproducing Kernel Hilbert Spaces in Probability and Statistics. Springer US, Boston, MA, p. 355
Bishop C., 2006, Pattern Recognition and Machine Learning. Springer, New York
Cannon C., 1973a, J. Quant. Spectrosc. Radiat. Transfer, 13, 627
Cannon C. J., 1973b, ApJ, 185, 621
Cockayne J., Oates C., Sullivan T., Girolami M., 2017, in Verdoolaege G., ed., AIP Conf. Proc., Am. Inst. Phys., New York, p. 060001
Cockayne J., Oates C. J., Sullivan T. J., Girolami M., 2019, SIAM Rev., 61, 756
Conrad P. R., Girolami M., Särkkä S., Stuart A., Zygalakis K., 2017, Stat. Comput., 27, 1065
Daubechies I., 1992, Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics, Philadelphia, PA
Decin L. et al., 2020, Science, 369, 1497
De Ceuster F., Homan W., Yates J., Decin L., Boyle P., Hetherington J., 2019, MNRAS, 492, 1812
De Ceuster F. et al., 2020, MNRAS, 499, 5194
de Mijolla D., Viti S., Holdship J., Manolopoulou I., Yates J., 2019, A&A, 630, A117
Diaconis P., 1988, in Gupta S. S., Berger J. O., eds, Statistical Decision Theory and Related Topics IV. Springer, New York, p. 163
Dykema P. G., Klein R. I., Castor J. I., 1996, ApJ, 457, 892
Fasshauer G. E., 1999, Adv. Comput. Math., 11, 139
Fornberg B., Flyer N., 2015, Acta Numerica, 24, 215
Gingold R. A., Monaghan J. J., 1977, MNRAS, 181, 375
Girolami M., Febrianto E., Yin G., Cirak F., 2021, Comput. Methods Appl. Mech. Eng., 375, 113533
Graepel T., 2003, in Fawcett T., Mishra N., eds, Proceedings, Twentieth International Conference on Machine Learning. The AAAI Press, Menlo Park, California, p. 234
Grassi T., Nauman F., Ramsey J. P., Bovino S., Picogna G., Ercolano B., 2021, Reducing the Complexity of Chemical Networks Via Interpretable Autoencoders, A&A. Forthcoming, https://doi.org/10.1051/0004-6361/202039956
Hennig P., Osborne M. A., Girolami M., 2015, Proc. R. Soc., 471, 20150142
Hennig P., Osborne M. A., Kersting H. P., 2022, Probabilistic Numerics. Cambridge Univ. Press, Cambridge
Holdship J., Viti S., Haworth T. J., Ilee J. D., 2021, A&A, 653, A76
Hubeny I., Mihalas D., 2014, Theory of Stellar Atmospheres. Princeton Univ. Press, Princeton
Kansa E., 1990a, Comput. Math. Appl., 19, 127
Kansa E., 1990b, Comput. Math. Appl., 19, 147
Kanschat G., Meinköhn E., Rannacher R., Wehrse R., von Waldenfels W., Cardall C. Y., 2009, Numerical Methods in Multidimensional Radiative Transfer. Springer, Berlin, Heidelberg
Kasim M. F. et al., 2022, Mach. Learn.: Sci. Technol., 3, 015013
Korčáková D., Kubát J., 2003, A&A, 401, 419
Lagaris I. E., Likas A., Fotiadis D. I., 1998, IEEE Trans. Neural Netw., 9, 987
Lagaris I. E., Likas A. C., Papageorgiou D. G., 2000, IEEE Trans. Neural Netw., 11, 1041
Lucy L. B., 1977, AJ, 82, 1013
McElroy D., Walsh C., Markwick A. J., Cordiner M. A., Smith K., Millar T. J., 2013, A&A, 550, A36
Meier D. L., 1999, ApJ, 518, 788
Metropolis N., Ulam S., 1949, J. Am. Stat. Assoc., 44, 335
Mihalas D., Weibel-Mihalas B., 1984, Foundations of Radiation Hydrodynamics. Oxford Univ. Press, Oxford
Mishra S., Molinaro R., 2021, J. Quant. Spectrosc. Radiat. Transfer, 270, 107705
Mishra S., Molinaro R., 2022, IMA J. Numer. Anal., 42, 981
Moens N., Sundqvist J. O., El Mellah I., Poniatowski L., Teunissen J., Keppens R., 2022, A&A, 657, A81
Noebauer U. M., Sim S. A., 2019, Living Rev. Comput. Astrophys., 5, 1
Oates C. J., Sullivan T. J., 2019, Stat. Comput., 29, 1335
Oberkampf W. L., Roy C. J., 2010, Verification and Validation in Scientific Computing. Cambridge Univ. Press, Cambridge
Owhadi H., 2017, SIAM Rev., 59, 99
Owhadi H., Scovel C., 2019, Operator-Adapted Wavelets, Fast Solvers, and Numerical Homogenization. Cambridge Univ. Press, Cambridge
Poincaré H., 1896, Calcul des probabilités. Georges Carré, Paris
Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 2007, Numerical Recipes: The Art of Scientific Computing, 3rd edn. Cambridge University Press, Cambridge, UK
Raissi M., Perdikaris P., Karniadakis G. E., 2019, J. Comput. Phys., 378, 686
Rasmussen C. E., Williams C. K. I., 2006, Gaussian Processes for Machine Learning. The MIT Press, Cambridge, MA, USA
Richling S., Meinköhn E., Kryzhevoi N., Kanschat G., 2001, A&A, 380, 776
Rybicki G. B., Hummer D. G., 1991, A&A, 245, 171
Schaback R., Wendland H., 2006, Acta Numerica, 15, 543
Shah A., Wilson A. G., Ghahramani Z., 2014, in Kaski S., Corander J., eds, Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS). Reykjavik, Cambridge, MA, USA, p. 877
Stevenson R., 2009, in Multiscale, Nonlinear and Adaptive Approximation: Dedicated to Wolfgang Dahmen on the Occasion of his 60th Birthday. Springer, Berlin, Heidelberg, p. 543
Van de Sande M., Millar T. J., 2019, ApJ, 873, 36
van den Boogaart K. G., 2001, in Proceedings of the IAMG2001. Cancun, Mexiko, p. 1, http://www.kgs.ku.edu/Conferences/IAMG//Sessions/E/Papers/boogaartpdf, last accessed date 05/08/2022
Verfürth R., 2013, A Posteriori Error Estimation Techniques for Finite Element Methods. Oxford Univ. Press, Oxford
Vetterli M., 2001, IEEE Signal Process. Mag., 18, 59
Xia C., Teunissen J., Mellah I. E., Chané E., Keppens R., 2018, ApJS, 234, 30

## APPENDIX A: MATHEMATICAL BACKGROUND

### A1 Equivalence between primal and dual formulation

To show the equivalence between the primal and dual formulation we have to prove equation (12), or equivalently,

$$\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}+\boldsymbol{\lambda}^{-2}\right)^{-1}-\left(\boldsymbol{\Phi}\boldsymbol{\lambda}^{2}\boldsymbol{\Phi}^{\mathrm{T}}+\boldsymbol{\sigma}^{2}\right)^{-1}\boldsymbol{\Phi}\boldsymbol{\lambda}^{2}=0. \quad \text{(A1)}$$

Using the Woodburry matrix identity, the second term expands as,

$$\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}\boldsymbol{\lambda}^{2}-\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}\left(\boldsymbol{\lambda}^{-2}+\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}\boldsymbol{\lambda}^{2}. \quad \text{(A2)}$$

Using (A2) in (A1), ignoring the overall factor, $\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}$, and isolating the terms with the inverse, it remains to show that,

$$\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}+\boldsymbol{\lambda}^{-2}\right)^{-1}\left(\mathbb{1}+\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}\boldsymbol{\lambda}^{2}\right)-\boldsymbol{\lambda}^{2}=0. \quad \text{(A3)}$$

Rewriting the second factor by extracting $\boldsymbol{\lambda}^2$ then yields

$$\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}+\boldsymbol{\lambda}^{-2}\right)^{-1}\left(\boldsymbol{\lambda}^{-2}+\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\boldsymbol{\Phi}\right)\boldsymbol{\lambda}^{2}-\boldsymbol{\lambda}^{2}=0 \quad \text{(A4)}$$

making it clear that the equality indeed holds and that the primal and dual solutions are thus equivalent.

### A2 Definitions

In this section, we summarize and explain some of the concepts from statistics that are used throughout the main text.

#### *A2.1 Expectation*

The expectation, $\mathbb{E}[X]$, of a random variable, $X$, is defined as the integral (or sum) over all possible values, $x$, that variable can take, weighted by its probability density, $p(x)$,

$$\mathbb{E}[X] \equiv \int \mathrm{d}x\; p(x)\, x. \quad \text{(A5)}$$

Although here, in our notation, we carefully distinguished between the random variable, $X$, and a specific realization of that variable, $x$, throughout this paper, we sometimes make the common slight abuse of notation by denoting both a random variable and its realizations with the same symbol.

#### *A2.2 Variance*

The variance, $\mathbb{V}[X]$, of a random variable, $X$, is defined as the expectation of the square difference between that variable and its expectation,

$$\mathbb{V}[X] \equiv \mathbb{E}\left[(X-\mathbb{E}[X])^2\right]. \quad \text{(A6)}$$

The variance can be interpreted as the expected square deviation from its expectation and thus quantifies the spread of the distribution of the random variable. Sometimes, the variance can be computed more conveniently as,

$$\mathbb{V}[X] = \mathbb{E}\left[X^2\right]-\mathbb{E}[X]^2, \quad \text{(A7)}$$

which follows from the definition (A6) by direct computation.

#### *A2.3 Covariance*

The covariance, $\mathrm{Cov}[X_i, X_j]$, between two random variables, $X_i$ and $X_j$, is defined as the expectation of the product of the differences of each variable with its expectation,

$$\mathrm{Cov}[X_i, X_j] \equiv \mathbb{E}\left[(X_i-\mathbb{E}[X_i])\left(X_j-\mathbb{E}[X_j]\right)\right]. \quad \text{(A8)}$$

The covariance of a random variable and itself equals its variance,

$$\mathrm{Cov}[X_i, X_i] = \mathbb{V}[X_i], \quad \text{(A9)}$$

which follows directly from the definitions (A6) and (A8).

#### *A2.4 Marginal probability*

The marginal probability distribution, $p(X_i)$, of a random variable, $X_i$, given the joint probability distribution, $p(X_i, X_j)$, with another random variable, $X_j$, is given by

$$p(X_i) \equiv \int \mathrm{d}X_j\; p(X_i, X_j), \quad \text{(A10)}$$

which amounts to integrating out the other random variable. Note the abuse of notation in using the random variable, $X_j$, to denote its observed value, $x_j$.

#### *A2.5 Conditional probability*

The conditional probability, $p(X_i|x_j)$, of a random variable, $X_i$, given the observation of the value, $x_j$, of another random variable, $X_j$, is given by

$$p(X_i|X_j) \equiv \frac{p(X_i, X_j)}{\int \mathrm{d}X_i\; p(X_i, X_j)} = \frac{p(X_i, X_j)}{p(X_j)}, \quad \text{(A11)}$$

which amounts to a re-scaling of the joint distribution, $p(X_i, X_j)$, with the marginal distribution $p(X_j)$. Note the abuse of notation in using the random variable, $X_j$, to denote its observed value, $x_j$.

#### *A2.6 Multivariate Gaussian or normal probability distribution*

A random vector variable, $\boldsymbol{X} \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$, follows a multivariate Gaussian or normal probability distribution with mean vector, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$, if its probability distribution is given by

$$p(X) \equiv \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\boldsymbol{X}-\mu)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\mu)\right). \quad \text{(A12)}$$

By direct computation, one can verify that the expectation, variance, and covariance of the components, $X_i$, of the multivariate Gaussian distributed vector variable $\boldsymbol{X} \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$, are respectively given by

$$\mathbb{E}[X_i] = \mu_i, \quad \text{(A13)}$$

$$\mathbb{V}[X_i] = \Sigma_{ii}, \quad \text{(A14)}$$

$$\mathrm{Cov}[X_i, X_j] = \Sigma_{ij}. \quad \text{(A15)}$$

The relations between a marginal and conditional (multivariate) Gaussian distributions are given in Appendix A3 and the relations for conditioning a (multivariate) Gaussian distribution are given in Appendix A4.

### A3 Marginal and conditional Gaussians

Given a (marginal) Gaussian distribution, $p(\boldsymbol{x})$, and a corresponding conditional Gaussian distribution, $p(\boldsymbol{y}\,|\,\boldsymbol{x})$, which are defined as

$$p(\boldsymbol{x}) = \mathcal{N}\left(\mu_{\boldsymbol{x}},\ \boldsymbol{\Sigma}_{\boldsymbol{x}}\right), \quad \text{(A16)}$$

$$p(\boldsymbol{y}\,|\,\boldsymbol{x}) = \mathcal{N}\left(A\boldsymbol{x}+\boldsymbol{b},\ \boldsymbol{\Sigma}_{\boldsymbol{y}|\boldsymbol{x}}\right), \quad \text{(A17)}$$

the other corresponding marginal distribution, $p(\boldsymbol{y})$, and the reverse conditional distribution, $p(\boldsymbol{x}\,|\,\boldsymbol{y})$, are given by

$$p(\boldsymbol{y}) \;=\; \mathcal{N}\left(A\mu_{\boldsymbol{x}} + \boldsymbol{b},\;\; \boldsymbol{\Sigma}_{y|x} \;+\; A\boldsymbol{\Sigma}_{\boldsymbol{x}}A^{\mathrm{T}}\right), \tag{A18}$$

$$p(\boldsymbol{x}\,|\,\boldsymbol{y}) \;=\; \mathcal{N}\Big(\boldsymbol{\Sigma}\left(A^{\mathrm{T}}\boldsymbol{\Sigma}_{y|x}^{-1}(\boldsymbol{y}-\boldsymbol{b}) \;+\; \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\mu_{\boldsymbol{x}}\right),\;\; \boldsymbol{\Sigma}\Big), \tag{A19}$$

in which we defined the covariance matrix,

$$\boldsymbol{\Sigma} \;\equiv\; \left(\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \;+\; A^{\mathrm{T}}\boldsymbol{\Sigma}_{y|x}^{-1}A\right)^{-1}. \tag{A20}$$

These relations can be derived by 'completing the square' in the distribution function and collecting the relevant terms, as described in detail, for instance, in Bishop (2006).

### A4 Conditioning a Gaussian

Consider a stochastic vector variable, $\boldsymbol{y}$, defined by two separate stochastic vector variables, $\boldsymbol{a}$ and $\boldsymbol{b}$, and assume that all components follow a (multivariate) Gaussian distribution, i.e.

$$\boldsymbol{y} \;=\; \begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} \;\sim\; \mathcal{N}\left(\begin{bmatrix}\mu_{\boldsymbol{a}}\\ \mu_{\boldsymbol{b}}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{aa}} & \boldsymbol{\Sigma}_{\boldsymbol{ab}}\\ \boldsymbol{\Sigma}_{\boldsymbol{bb}} & \boldsymbol{\Sigma}_{\boldsymbol{bb}}\end{bmatrix}\right), \tag{A21}$$

in which, $\boldsymbol{\mu}_{\boldsymbol{a}}$ and $\boldsymbol{\mu}_{\boldsymbol{b}}$ are the mean vectors and the matrices $\boldsymbol{\Sigma}_{\boldsymbol{aa}}$, $\boldsymbol{\Sigma}_{\boldsymbol{ab}} = \boldsymbol{\Sigma}_{\boldsymbol{ba}}^{\mathrm{T}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{bb}}$, together form the covariance matrix. Now, we can ask what the resulting distribution of $\boldsymbol{a}$ would be, given prior knowledge about the value for $\boldsymbol{b}$. Fixing the value of $\boldsymbol{b}$ again yields a multivariate Gaussian distribution,

$$p\left(\boldsymbol{a}\,|\,\boldsymbol{b}\right) \;=\; \mathcal{N}\big(\mu_{\boldsymbol{a}|\boldsymbol{b}},\,\boldsymbol{\Sigma}_{\boldsymbol{a}|\boldsymbol{b}}\big), \tag{A22}$$

in which the conditioned mean and variance are given by

$$\boldsymbol{\mu}_{\boldsymbol{a}|\boldsymbol{b}} \;=\; \mu_{\boldsymbol{a}} \;+\; \boldsymbol{\Sigma}_{\boldsymbol{ab}}\,\boldsymbol{\Sigma}_{\boldsymbol{bb}}^{-1}\left(\boldsymbol{b}-\boldsymbol{\mu}_{\boldsymbol{b}}\right), \tag{A23}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{a}|\boldsymbol{b}} \;=\; \boldsymbol{\Sigma}_{\boldsymbol{aa}} \;-\; \boldsymbol{\Sigma}_{\boldsymbol{ab}}\,\boldsymbol{\Sigma}_{\boldsymbol{bb}}^{-1}\,\boldsymbol{\Sigma}_{\boldsymbol{ba}}. \tag{A24}$$

These relations can be derived by 'completing the square' in the distribution function and collecting the relevant terms, as described in detail, for instance, in Bishop (2006).

Note that without correlation between $\boldsymbol{a}$ and $\boldsymbol{b}$, i.e. when $\boldsymbol{\Sigma}_{\boldsymbol{ab}} = \boldsymbol{\Sigma}_{\boldsymbol{ba}}^{\mathrm{T}} = 0$, the prior knowledge about $\boldsymbol{b}$ will not affect the distribution of $\boldsymbol{a}$, which is in line with expectations.

### A5 RKHS bound on the uncertainty

Let $\mathcal{H}$ denote the reproducing kernel Hilbert space (RKHS) of the kernel defined in (7), with an associated inner product, $\langle\rangle_{\mathcal{H}}$, and norm $\|\cdot\|_{\mathcal{H}} \equiv \sqrt{\langle\cdot,\cdot\rangle_{\mathcal{H}}}$. The defining properties of an RKHS with reproducing kernel, $k$, are that, $k(x,\cdot) \in \mathcal{H}$, and that,

$$\forall f \in \mathcal{H}:\; \langle f, k(x,\cdot)\rangle_{\mathcal{H}} \;=\; f(x), \tag{A25}$$

i.e. an inner product with the kernel around $x$ corresponds to function evaluation in $x$ (see e.g. Berlinet & Thomas-Agnan 2004, for a comprehensive introduction). The latter is known as the reproducing property and it is the key to derive the bound given in equation (48). If we now consider the projection $Pf \in \mathcal{H}$ of the function, $f$, in the RKHS, $\mathcal{H}$, the reproducing property (A25) implies that

$$Pf(x) \;=\; \langle Pf, k(x,\cdot)\rangle_{\mathcal{H}}\,. \tag{A26}$$

By definition of the data, we also have that $Pf(\boldsymbol{x}) = \mathbf{y}$, such that

$$\mathbf{y} \;=\; \langle Pf, k(x,\cdot)\rangle_{\mathcal{H}}\,. \tag{A27}$$

Substituting this in (31), and defining $\mathbf{K} \equiv k(x,x) + \boldsymbol{\sigma}^2$, yields

$$\tilde{f}(x) \;=\; k(x,\boldsymbol{x})\,\mathbf{K}^{-1}\,\langle Pf, k(x,\cdot)\rangle_{\mathcal{H}}\,, \tag{A28}$$

such that, in combination with equation (A26), we find that

$$\begin{aligned}\left|Pf(x) - \tilde{f}(x)\right| &= \left|\left\langle Pf,\; k(\cdot,x) - k(x,\boldsymbol{x})\,\mathbf{K}^{-1}k(\cdot,x)\right\rangle_{\mathcal{H}}\right|\\ &\leq \|Pf\|_{\mathcal{H}}\,\left\|k(x,\cdot) - k(x,\boldsymbol{x})\,\mathbf{K}^{-1}k(x,\cdot)\right\|_{\mathcal{H}},\end{aligned} \tag{A29}$$

where in the last step, we used the Cauchy–Schwarz inequality. Considering the square of the last factor, we find

$$\begin{aligned}\left\|k(x,\cdot) - k(x,\boldsymbol{x})\,\mathbf{K}^{-1}k(x,\cdot)\right\|_{\mathcal{H}}^2 &= k(x,x) \;-\; 2k(x,\boldsymbol{x})\,\mathbf{K}^{-1}k(x,x)\\ &\quad + k(x,\boldsymbol{x})\,\mathbf{K}^{-1}k(x,x)\,\mathbf{K}^{-1}k(x,x)\\ &= k(x,x) \;-\; k(x,\boldsymbol{x})\,\mathbf{K}^{-1}k(x,x)\\ &\quad -k(x,\boldsymbol{x})\,\mathbf{K}^{-1}\left(\mathbf{K} - k(x,x)\right)\mathbf{K}^{-1}k(x,x)\\ &= \tilde{\varepsilon}^2(x) \;-\; k(x,\boldsymbol{x})\,\mathbf{K}^{-1}\boldsymbol{\sigma}^2\,\mathbf{K}^{-1}k(x,x),\end{aligned} \tag{A30}$$

where in the last equality we used equation (32). Since the second term can be viewed as (minus) the square of the Euclidean norm of the vector, $\boldsymbol{\sigma}\,\mathbf{K}^{-1}k(x,x)$, it will always be negative, such that,

$$\left\|k(x,\cdot) - k(x,\boldsymbol{x})\,\mathbf{K}^{-1}k(x,\cdot)\right\|_{\mathcal{H}} \;\leq\; \tilde{\varepsilon}(x). \tag{A31}$$

Note that in the limit of perfect data, i.e. $\boldsymbol{\sigma} \to \mathbf{0}$, the above inequality becomes an equality. Substituting this in equation (A29), we obtain the desired bound on the local error,

$$\left|Pf(x) - \tilde{f}(x)\right| \;\leq\; \|Pf\|_{\mathcal{H}}\;\tilde{\varepsilon}(x). \tag{A32}$$

It should be emphasized that this only bounds the absolute difference between the approximation and the projection of the true solution in the RKHS, not the absolute difference between the approximation and the true solution itself. Therefore, the strength of this bound crucially depends on the RKHS, and thus on the particular kernel, or equivalently, on the particular set of basis functions that is used.

### A6 Equivalent kernel for the method of characteristics

Given a linear PDE, and given the corresponding Green's function, $G$, for the differential operator, one can construct a kernel,

$$k(z,s) \;=\; \int_{-\infty}^{+\infty}\mathrm{d}s'\int_{-\infty}^{+\infty}\mathrm{d}z'\;\kappa(s',z')\;G(z',z)\;G(s',s), \tag{A33}$$

based on another kernel, $\kappa$. For later convenience, we define a new function, $g$, that, using the Green's functions, can be expressed as

$$g(s,z) \;\equiv\; \mathscr{L}_2 k(z,s) \;=\; \int_{-\infty}^{+\infty}\mathrm{d}z'\;\kappa(s,z')\;G(z',z), \tag{A34}$$

$$g(z,s) \;\equiv\; \mathscr{L}_1 k(z,s) \;=\; \int_{-\infty}^{+\infty}\mathrm{d}s'\;\kappa(s',z)\;G(s',s), \tag{A35}$$

in which the subscript on the differential operator, $\mathscr{L}$, indicates whether it acts on the first or second argument. Note that both definitions are consistent, since $k$ is symmetric in its arguments. Using the Green's functions again, one can derive,

$$\mathscr{L}_1\mathscr{L}_2 k(z,s) \;=\; \mathscr{L}_1 g(s,z) \;=\; \kappa(s,z), \tag{A36}$$

$$\mathscr{L}_2\mathscr{L}_1 k(z,s) \;=\; \mathscr{L}_2 g(z,s) \;=\; \kappa(s,z). \tag{A37}$$

When solving the PDE as a Bayesian linear regression problem, the corresponding covariance matrix of the joint distribution, reads

$$\begin{pmatrix}\mathscr{L}_1\mathscr{L}_2 k(\boldsymbol{a},\boldsymbol{a}) & \mathscr{L}_1\mathscr{B}_2 k(\boldsymbol{a},\boldsymbol{b}) & \mathscr{L}_1 k(\boldsymbol{a},s)\\ \mathscr{B}_1\mathscr{L}_2 k(\boldsymbol{b},\boldsymbol{a}) & \mathscr{B}_1\mathscr{B}_2 k(\boldsymbol{b},\boldsymbol{b}) & \mathscr{B}_1 k(\boldsymbol{b},s)\\ \mathscr{L}_2 k(s,\boldsymbol{a}) & \mathscr{B}_2 k(s,\boldsymbol{b}) & k(s,s)\end{pmatrix} \tag{A38}$$

and can be simplified using the definitions above to yield

$$\begin{pmatrix}\kappa(\boldsymbol{a},\boldsymbol{a}) & \mathscr{B}_1 g(\boldsymbol{a},\boldsymbol{b}) & g(\boldsymbol{a},s)\\ \mathscr{B}_1 g(\boldsymbol{a},\boldsymbol{b}) & \mathscr{B}_1\mathscr{B}_2 k(\boldsymbol{b},\boldsymbol{b}) & \mathscr{B}_1 k(\boldsymbol{b},s)\\ g(\boldsymbol{a},s) & \mathscr{B}_2 k(s,\boldsymbol{b}) & k(s,s)\end{pmatrix}. \tag{A39}$$

The requirement that this matrix is positive semi-definite for all Green's functions, $G$, can be reduced to the condition that

$$\kappa(\boldsymbol{a},\, s)^{\mathrm{T}}\, \kappa(\boldsymbol{a},\, \boldsymbol{a})^{-1}\, \kappa(\boldsymbol{a},\, z) \;\leq\; \kappa(s,\, z), \tag{A40}$$

holds for all $s, z \in D$, which is equivalent to the condition that $\kappa$ is a positive semi-definite kernel, as expected.

In the method of characteristics (Section 3.2), we considered the special case where the second kernel, $\kappa$, has the additional property that it cannot correlate the regions $s > s_0$ and $s < s_0$, i.e.

$$\begin{aligned}\kappa(z, s) \equiv\; & \Theta(s_0 - z)\Theta(s_0 - s)\kappa(z, s) \\ & + \Theta(z - s_0)\Theta(s - s_0)\kappa(z, s).\end{aligned} \tag{A41}$$

Using the Green's function from the radiative transfer equation, this implies, for $z \geq b$ and $s \geq b$, that

$$\begin{aligned}k(z, s) \;=\; & \int_b^z \mathrm{d}z' \int_b^s \mathrm{d}s'\, \kappa(z',\, s')\, e^{-\tau(z',z)}\, e^{-\tau(s',s)} \\ & +\; k(b, b)\, e^{-\tau(b,z)}\, e^{-\tau(b,s)}.\end{aligned} \tag{A42}$$

Similarly, this implies, for $z \geq b$ and $s \geq b$, that

$$g(s, z) \;=\; \int_b^z \mathrm{d}z'\, \kappa(s, z')\, e^{-\tau(z',z)}, \tag{A43}$$

and in particular that for $s \geq b$, we have that $g(s, b) = 0$. As a result, the inverted matrix in equations (33) and (34) reduces to

$$\begin{aligned}&\begin{pmatrix} \mathscr{L}_1\mathscr{L}_2 k(\boldsymbol{a}, \boldsymbol{a}) + \boldsymbol{\sigma}_{\mathrm{L}}^2 & \mathscr{L}_1\mathscr{B}_2 k(\boldsymbol{a}, b) \\ \mathscr{B}_1\mathscr{L}_2 k(b, \boldsymbol{a}) & \mathscr{B}_1\mathscr{B}_2 k(b, b) + \sigma_{\mathrm{B}}^2 \end{pmatrix} \\ &\quad = \begin{pmatrix} \kappa(\boldsymbol{a}, \boldsymbol{a}) + \boldsymbol{\sigma}_{\mathrm{L}}^2 & g(\boldsymbol{a}, b) \\ g(\boldsymbol{a}, b)^{\mathrm{T}} & k(b, b) + \sigma_{\mathrm{B}}^2 \end{pmatrix} \\ &\quad = \begin{pmatrix} \kappa(\boldsymbol{a}, \boldsymbol{a}) + \boldsymbol{\sigma}_{\mathrm{L}}^2 & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & k(b, b) + \sigma_{\mathrm{B}}^2 \end{pmatrix}.\end{aligned} \tag{A44}$$

Define the matrix $\mathbf{K} \equiv \kappa(\boldsymbol{a}, \boldsymbol{a}) + \boldsymbol{\sigma}_{\mathrm{L}}^2$, and let us assume that there is no uncertainty on the boundary condition, i.e. $\sigma_{\mathrm{B}} = 0$. The function approximation in the dual formulation then reads

$$\begin{aligned}\tilde{f}_{\mathrm{dual}}(s) \;&=\; \begin{pmatrix} \boldsymbol{\eta} \\ I_0 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \mathbf{K} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & k(b, b) \end{pmatrix}^{-1} \begin{pmatrix} g(\boldsymbol{a}, s) \\ k(b, s) \end{pmatrix} \\ &= I_0\, e^{-\tau(b,s)} + \boldsymbol{\eta}^{\mathrm{T}}\mathbf{K}^{-1} \textstyle\int_b^s \mathrm{d}s'\, \kappa(\boldsymbol{a}, s') e^{-\tau(s',s)}.\end{aligned} \tag{A45}$$

We clearly recognize the result from the method of characteristics. Similarly, the corresponding variance reads

$$\begin{aligned}\tilde{\sigma}^2_{\mathrm{dual}}(s) = k(s, s) - & \begin{pmatrix} g(\boldsymbol{a}, s) \\ k(b, s) \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \mathbf{K} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & k(b, b) \end{pmatrix}^{-1} \begin{pmatrix} g(\boldsymbol{a}, s) \\ k(b, s) \end{pmatrix} \\ &= \textstyle\int_b^s \mathrm{d}s' \int_b^s \mathrm{d}z'\, e^{-\tau(s',\, s)} e^{-\tau(z',\, s)} \\ &\quad \times \left(\kappa(s',\, z') - \kappa(\boldsymbol{a},\, s')^{\mathrm{T}}\mathbf{K}^{-1}\kappa(\boldsymbol{a},\, z')\right).\end{aligned} \tag{A46}$$

In the parentheses, we recognize the conditioned variance (32) that stems from the interpolation of the emissivity (70).

### A7 The law of total expectation

Given two random variables, $X$ and $Y$, in the same probability space, the law of total expectation states that

$$\mathbb{E}[X] \;=\; \mathbb{E}_Y\big[\mathbb{E}[X|Y]\big], \tag{A47}$$

i.e. the expectation of $X$ is the same as the expectation over $Y$ of the conditional expectation of $X$ given $Y$. A sketch for a proof can be derived from the following equalities,

$$\begin{aligned}\mathbb{E}_Y\big[\mathbb{E}[X|Y]\big] \;&=\; \int \mathrm{d}Y\; p(Y) \int \mathrm{d}X\; p(X|Y)\, X \\ &= \int \mathrm{d}Y \int \mathrm{d}X\; p(X, Y)\, X \\ &= \int \mathrm{d}X\; p(X)\, X \\ &= \mathbb{E}[X],\end{aligned} \tag{A48}$$

where in the first equality we used the definition of the expectation, in the second we used the conditional probability, and in the third we used the marginal probability.

### A8 The law of total total variance

Given two random variables, $X$ and $Y$, in the same probability space, the law of total variance states that,

$$\mathbb{V}[X] \;=\; \mathbb{E}_Y\big[\mathbb{V}[X|Y]\big] \;+\; \mathbb{V}_Y\big[\mathbb{E}[X|Y]\big], \tag{A49}$$

i.e. the variance of $X$ is the sum of the expected conditional variance and the variance of the conditional expectation. This follows directly form the law of total expectation (A47). Using the law of total expectation in equation (A7) yields

$$\begin{aligned}\mathbb{V}[X] =\; & \mathbb{E}_Y\Big[\mathbb{E}\big[X^2|Y\big]\Big] \;-\; \mathbb{E}_Y\big[\mathbb{E}[X|Y]\big]^2 \\ &= \mathbb{E}_Y\Big[\mathbb{V}\big[X|Y\big] \;+\; \mathbb{E}\big[X|Y\big]^2\Big] \;-\; \mathbb{E}_Y\big[\mathbb{E}[X|Y]\big]^2 \\ &= \mathbb{E}_Y\Big[\mathbb{V}\big[X|Y\big]\Big] \;+\; \mathbb{E}_Y\Big[\mathbb{E}\big[X|Y\big]^2\Big] \;-\; \mathbb{E}_Y\big[\mathbb{E}[X|Y]\big]^2 \\ &= \mathbb{E}_Y\big[\mathbb{V}[X|Y]\big] \;+\; \mathbb{V}_Y\big[\mathbb{E}[X|Y]\big],\end{aligned} \tag{A50}$$

where in the second and fourth equality we used equation (A7) and in the third equality we used the linearity of the expectation.

### A9 Expectations of optical depth

We compute the expectations with respect to the Gaussian-distributed optical depth to obtain the total expectation and variance in equations (87) and (88). Since the optical depth always appears in an exponential, we are interested in,

$$\begin{aligned}\mathbb{E}_\tau\left[e^{-\tau(z,s)}\right] \;&=\; \int \mathrm{d}\tau(z, s)\; p\big(\tau(z, s)\big)\, e^{-\tau(z,s)} \\ &= \exp\left(-\tilde{\tau}(z, s) \;+\; \frac{1}{2}\,\tilde{\varepsilon}^2_\tau(z, s)\right),\end{aligned} \tag{A51}$$

in which we used that the optical depth is Gaussian distributed, as in equation (84). In the exponent, we recognize what we defined as the effective optical depth in equation (89). Using the linearity of the expectation, this immediately yields the result in equation (87).

Similarly, for the variance, we also require,

$$\begin{aligned}\mathbb{E}_\tau\left[e^{-\tau(z',s)} e^{-\tau(s',s)}\right] = \mathbb{E}_\tau\left[e^{-\tau(z',s)}\right] \mathbb{E}_\tau\left[e^{-\tau(s',s)}\right] \\ +\mathrm{Cov}\left[e^{-\tau(z',s)},\; e^{-\tau(s',s)}\right].\end{aligned} \tag{A52}$$

Since the optical depth is Gaussian distributed, the exponential of minus the optical depth will follow a log-normal distribution. The

covariance of this log-normal distribution is given by

$$\begin{aligned}&\mathrm{Cov}\left[e^{-\tau(z',s)},\ e^{-\tau(s',s)}\right]\\&\quad=\left(\exp\left(\mathrm{Cov}\left[-\tau(z',s),\ -\tau(s',s)\right]\right)\ -\ 1\right)\\&\quad\times\exp\left(-\tilde{\tau}(z',s)-\tilde{\tau}(s',s)\ +\ \frac{1}{2}\left(\tilde{\varepsilon}_\tau^2(z',s)+\tilde{\varepsilon}_\tau^2(s',s)\right)\right),\end{aligned}\tag{A53}$$

such that we can write the required expectation as

$$\begin{aligned}&\mathbb{E}_\tau\left[e^{-\tau(z',s)}e^{-\tau(s',s)}\right]=\exp\left(\mathrm{Cov}\left[-\tau(z',s),\ -\tau(s',s)\right]\right)\\&\quad\times\exp\left(-\tilde{\tau}(z',s)-\tilde{\tau}(s',s)\ +\ \frac{1}{2}\left(\tilde{\varepsilon}_\tau^2(z',s)+\tilde{\varepsilon}_\tau^2(s',s)\right)\right).\end{aligned}\tag{A54}$$

The covariance of the optical depths can easily be derived from their joint Gaussian distribution, which yields

$$\begin{aligned}&\mathrm{Cov}\left[-\tau(z',s),\ -\tau(s',s)\right]\\&\quad=\int_{z'}^{s}\mathrm{d}z''\int_{s'}^{s}\mathrm{d}s''\left(\kappa(s'',z'')-\kappa(\boldsymbol{a},s'')^{\mathrm{T}}\mathbf{K}_\chi^{-1}\kappa(\boldsymbol{a},z'')\right)\\&\quad=\frac{1}{2}\left(\tilde{\varepsilon}_\tau^2(z',s)\ +\ \tilde{\varepsilon}_\tau^2(s',s)\ -\ \tilde{\varepsilon}_\tau^2(s',z')\right),\end{aligned}\tag{A55}$$

where the last equality can be derived by subdividing the (2D) domain of integration. Using the second effective optical depth variable (90) to simplify notation, equation (A54) can be written as

$$\mathbb{E}_\tau\left[e^{-\tau(z',s)}e^{-\tau(s',s)}\right]\ =\ e^{-\overline{\tau}(z',s)}e^{-\overline{\tau}(s',s)}e^{-\frac{1}{2}\tilde{\varepsilon}_\tau^2(s',z')}.\tag{A56}$$

Using the linearity of the expectation, we thus find

$$\begin{aligned}\mathbb{E}_\tau\left[\tilde{\varepsilon}_I^2\right]\ =&\int_{s_0}^{s}\mathrm{d}s'\int_{s_0}^{s}\mathrm{d}z'e^{-\overline{\tau}(s',s)}e^{-\overline{\tau}(z',s)}e^{-\frac{1}{2}\tilde{\varepsilon}_\tau^2(s',z')}\\&\times\left(\kappa(s',\ z')-\kappa(\boldsymbol{a},\ s')^{\mathrm{T}}\mathbf{K}_\eta^{-1}\kappa(\boldsymbol{a},\ z')\right).\end{aligned}\tag{A57}$$

Furthermore, using the same relations, we can find that

$$\begin{aligned}\mathbb{E}_\tau\left[\tilde{I}^2\right]\ =&\ I_0^2\,e^{-2\overline{\tau}(s_0,s)}\\&+2\,I_0\,e^{-\overline{\tau}(s_0,s)}\int_{s_0}^{s}\mathrm{d}s'H(s')\,e^{-\frac{1}{2}\tilde{\varepsilon}_\tau^2(s_0,s')}\\&+\int_{s_0}^{s}\mathrm{d}z'H(z')\int_{s_0}^{s}\mathrm{d}s'H(s')\,e^{-\frac{1}{2}\tilde{\varepsilon}_\tau^2(z',s')}\end{aligned}\tag{A58}$$

where we defined $H(s')\equiv\boldsymbol{\eta}^{\mathrm{T}}\,\mathbf{K}_\eta^{-1}\kappa(\boldsymbol{a},s')\ e^{-\overline{\tau}(s',s)}$, to simplify notation. In practice, equations (A57) and (A58) are difficult to work with due to their dependence on the cross term, $\tilde{\varepsilon}_\tau^2(z',s')$. However, since this is a positive quantity, we can define an upper bound by removing it, such that equation (A57) simplifies to

$$\begin{aligned}\mathbb{E}_\tau\left[\tilde{\varepsilon}_I^2\right]\ \leq&\int_{s_0}^{s}\mathrm{d}s'\int_{s_0}^{s}\mathrm{d}z'e^{-\overline{\tau}(s',s)}e^{-\overline{\tau}(z',s)}\\&\times\left(\kappa(s',\ z')-\kappa(\boldsymbol{a},\ s')^{\mathrm{T}}\mathbf{K}_\eta^{-1}\kappa(\boldsymbol{a},\ z')\right),\end{aligned}\tag{A59}$$

and, furthermore, equation (A58) simplifies to

$$\mathbb{E}_\tau\left[\tilde{I}^2\right]\ \leq\ \bar{I}^2.\tag{A60}$$

where, in analogy with equation (A47), we defined

$$\bar{I}(s)\ \equiv\ I_0\,e^{-\overline{\tau}(s_0,s)}\ +\ \boldsymbol{\eta}^{\mathrm{T}}\,\mathbf{K}_\eta^{-1}\int_{s_0}^{s}\mathrm{d}s'\,\kappa(\boldsymbol{a},s')\,e^{-\overline{\tau}(s',s)}\tag{A61}$$

Combining all these results yields the practical inequality (88).

This paper has been typeset from a TEX/LATEX file prepared by the author.