Journal of the Royal Statistical Society Series B: Statistical Methodology, 2025, **87**, 457–479
https://doi.org/10.1093/jrsssb/qkae098
Advance access publication 26 December 2024
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology

Probabilistic Richardson extrapolation

Chris. J. Oates[1], Toni Karvonen[2,3], Aretha L. Teckentrup[4], Marina Strocchi[5,6] and Steven A. Niederer[5,6]

[1]School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne, UK
[2]School of Engineering Sciences, Lappeenranta–Lahti University of Technology LUT, Lappeenranta, Finland
[3]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
[4]School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh, Edinburgh, UK
[5]National Heart and Lung Institute, Imperial College London, London, UK
[6]School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK

Address for correspondence: Chris. J. Oates, School of Mathematics, Statistics & Physics, Newcastle University, Newcastle upon Tyne NE1 7RU, UK. Email: chris.oates@ncl.ac.uk

Abstract

For over a century, extrapolation methods have provided a powerful tool to improve the convergence order of a numerical method. However, these tools are not well-suited to modern computer codes, where multiple continua are discretized and convergence orders are not easily analysed. To address this challenge, we present a probabilistic perspective on Richardson extrapolation, a point of view that unifies classical extrapolation methods with modern multi-fidelity modelling, and handles uncertain convergence orders by allowing these to be statistically estimated. The approach is developed using Gaussian processes, leading to *Gauss–Richardson Extrapolation*. Conditions are established under which extrapolation using the conditional mean achieves a polynomial (or even an exponential) speed-up compared to the original numerical method. Further, the probabilistic formulation unlocks the possibility of experimental design, casting the selection of fidelities as a continuous optimization problem, which can then be (approximately) solved. A case study involving a computational cardiac model demonstrates that practical gains in accuracy can be achieved using the GRE method.

Keywords: Bayesian statistics, Gaussian process, multi-fidelity modelling, reproducing kernel, uncertainty quantification

1 Introduction

Testing of hypotheses underpins the scientific method, and increasingly these hypotheses are model-based. Deterministic or stochastic mathematical models are routinely used to represent mechanisms hypothesized to govern diverse phenomena, such as aerodynamics or electrochemical regulation of the human heart. In these cases, critical scientific enquiry demands a comparison of the model against a real-world dataset. The practical challenge is twofold; to simulate from the mathematical model, and to obtain a real-world dataset. Here, we focus on the first challenge—simulating from the model—which can be arbitrarily difficult depending on the complexity of the model. For example, simulating a single cycle of a jet engine to an acceptable numerical precision routinely requires 10^6 core hours (Arroyo et al., 2021), while accurate simulation from the cardiac models that we consider later in this paper at steady state requires 10^4 core hours in total (Strocchi et al., 2023). To drive progress in these and many other diverse scientific domains, there is an urgent need for statistical and computational methodology that can mitigate the high cost of accurately simulating from a mathematical model.

Abstractly, we enumerate all of the *discretization parameters* involved in approximate simulation from the mathematical model using scalars $\mathbf{x} = (x_1, \ldots, x_d)$, such that each component of $\mathbf{x}$

Received: January 22, 2024. **Revised:** September 2, 2024. **Accepted:** September 16, 2024

controls the error due to a particular aspect of discretization; for example, x_1 could be a time-step size, x_2 could be the width of a spatial mesh, and x_3 could be an error tolerance for an adaptive numerical method. The principal requirement is that the ideal mathematical model corresponds to the limit $\mathbf{x} \to \mathbf{0}$, where no discretization is performed. Given a value of $\mathbf{x}$, we denote as $f(\mathbf{x})$ the associated numerical approximation to the continuum quantity $f(\mathbf{0})$ from the mathematical model. The computational cost of such an evaluation will be denoted $c(\mathbf{x})$, with $c(\mathbf{0}) = \infty$ being typical. The computational challenge addressed in this paper is to produce an accurate approximation to $f(\mathbf{0})$, based on a dataset of simulations $\{f(\mathbf{x}_j)\}$, where $\{\mathbf{x}_j\} \subset (0, \infty)^d$, such that the computational cost of obtaining $\{f(\mathbf{x}_j)\}$ remains within a prescribed budget. For this initial discussion, we focus on scalar-valued model output, but we generalize to multivariate and infinite-dimensional model output in Section 2.9.

Several solutions have been proposed to perform approximate simulation at a reduced cost. In what follows, it is useful to draw a distinction between *extrapolation methods*, applicable to the situation, where a mathematical model is discretized for simulation and numerical analysis of the discretization error can be performed, and *modern solutions* that are typically applied to 'black box' computer codes for which numerical analysis is impractical.

Extrapolation Methods. A unified presentation of extrapolation methods, that includes the most widely used algorithms, is provided by the so-called *E-algorithm* (see the survey of Brezinski, 1989). The starting point is a (real-valued) convergent sequence, which in our setting, we interpret as a sequence of numerical approximations $(f(\mathbf{x}_m))_{m\in\mathbb{N}}$, where $\mathbf{x}_m$ is a vector of discretization parameters controlling the error in approximating the mathematical model, while $f(\mathbf{0})$ represents the continuum quantity of interest. The E-algorithm posits an *ansatz* that

$$f(\mathbf{x}_m) = f(\mathbf{0}) + a_1 g_1(m) + \ldots + a_{n-1} g_{n-1}(m) \tag{1}$$

for some unknown $a_1, \ldots, a_{n-1} \in \mathbb{R}$, some known functions $g_i : \mathbb{N} \to \mathbb{R}$, and all $m \in \mathbb{N}$. Then, instantiating (1) for $m, m+1, \ldots, m+n-1$, we may solve for the unknown $a_1, \ldots, a_{n-1}$ and $f(\mathbf{0})$ in terms of the n values $f(\mathbf{x}_m), \ldots, f(\mathbf{x}_{m+n-1})$. Indeed, solving this linear system for $f(\mathbf{0})$ leads to the estimator

$$S_m := S(f(\mathbf{x}_m), \ldots, f(\mathbf{x}_{m+n-1})) = \frac{\begin{vmatrix} f(\mathbf{x}_m) & \ldots & f(\mathbf{x}_{m+n-1}) \\ g_1(m) & \ldots & g_1(m+n-1) \\ \vdots & & \vdots \\ g_n(m) & \ldots & g_n(m+n-1) \end{vmatrix}}{\begin{vmatrix} 1 & \ldots & 1 \\ g_1(m) & \ldots & g_1(m+n-1) \\ \vdots & & \vdots \\ g_n(m) & \ldots & g_n(m+n-1) \end{vmatrix}}. \tag{2}$$

Under appropriate assumptions, the sequence $(S_m)_{m\in\mathbb{N}}$ constructed based on $(f(\mathbf{x}_m))_{m\in\mathbb{N}}$ as in (2) not only has the same limit, $f(\mathbf{0})$, but also converges to that limit faster in the sense that $\lim_{m\to\infty}(S_m - f(\mathbf{0}))/(f(\mathbf{x}_m) - f(\mathbf{0})) = 0$; for precise statements, see Chapter 2 of Brezinski and Zaglia (2013).

The principal classes of extrapolation methods concern either the case of a single discretization parameter x_m, or they maintain ambivalence about $\mathbf{x}_m$ by operating only on the values of the sequence $(f(\mathbf{x}_m))_{m\in\mathbb{N}}$. In either case, different extrapolation methods correspond to different basis functions g_i in (1). *Richardson extrapolation* corresponds to $g_i(m) = x_m^i$, in which case (1) is recognized as polynomial extrapolation to the origin (Richardson, 1911; Richardson & Gaunt, 1927). The existence of a Taylor expansion of f at the origin is sufficient to guarantee a polynomial-rate convergence acceleration using Richardson's method. Other examples of extrapolation methods include *Shanks' transformation* $g_i(m) = f(\mathbf{x}_{m+i}) - f(\mathbf{x}_{m+i-1})$ (Shanks, 1955), the *Germain–Bonne transformation* $g_i(m) = (f(\mathbf{x}_{m+1}) - f(\mathbf{x}_m))^i$ (Germain-Bonne, 1990), and

Thiele's rational extrapolation method $g_i(m) = x_m^i$, $g_{i+p}(m) = f(x_m)x_m^i$ for $i = 1, \ldots, p$, $n = 2p + 1$ (Bulirsch & Stoer, 1964; Larkin, 1967; Thiele, 1909). A careful numerical analysis of f is usually required to determine when a particular extrapolation method can be applied. To the best of our knowledge, ideas from statistics and uncertainty quantification do not feature prominently, if at all, in the literature on extrapolation methods. In addition, the question of how best to construct the sequence $(\mathbf{x}_m)_{m\in\mathbb{N}}$ under a constraint on the overall computational budget does not appear to have been systematically addressed. Further background can be found in the book-level treatment of Sidi (2003) and Brezinski and Zaglia (2013).

Though rather classical, extrapolation methods continue to find new and useful applications, including in optimal transport (Chizat et al., 2020), regularization and training of machine learning models (Bach, 2021), and sampling with Markov chain Monte Carlo (Durmus et al., 2016).

Modern Solutions. If the mathematical model additionally involves one or more degrees of freedom $\boldsymbol{\theta}$, numerical approximations $f_{\boldsymbol{\theta}}(\mathbf{x})$ are often required across a range of values for $\boldsymbol{\theta}$ to identify configurations that are consistent with observations from the real world. Since the introduction of additional degrees of freedom further complicates numerical analysis, this setting has motivated the development of black box methods that can be applied in situations, where numerical analysis is impractical. Among these, *emulation* and *multi-fidelity modelling* (MFM) are arguably the most prominent.

In *emulation*, one attempts to approximate the map $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}}(\mathbf{x}_{\text{hi-fi}})$, where the discretization parameters $\mathbf{x}_{\text{hi-fi}}$ are typically fixed and correspond to a suitably high fidelity (hi-fi) model. This enables the prediction of computer code output at values of $\boldsymbol{\theta}$ for which simulation was not performed (Sacks et al., 1989). A variety of sophisticated techniques have been developed to identify an appropriate basis or subspace in which an emulator can be constructed, such as *reduced order modelling* (Lucia et al., 2004). One drawback of emulation is that it can be *data hungry*; in applications for which it is only possible to perform a small number n of simulations, and for which insight from numerical analysis is unavailable, one usually cannot expect to obtain high-quality predictions. A second drawback is that emulation treats the discretized model $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}}(\mathbf{x}_{\text{hi-fi}})$ as the target, whereas in reality, the continuum mathematical model $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}}(0)$ is of principal interest.

A partial solution to the drawbacks of emulation is MFM, in which one supplements a small number of simulations from the hi-fi model $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}}(\mathbf{x}_{\text{hi-fi}})$ with a larger number of simulations from one or more cheaper low fidelity (lo-fi) models $\boldsymbol{\theta} \mapsto f_{\boldsymbol{\theta}}(\mathbf{x}_{\text{lo-fi}})$ (Peherstorfer et al., 2018). Lo-fi models can sometimes be obtained using coarse-grid approximations, early stopping of iterative algorithms, or linearization (Piperni et al., 2013). Alternatively, lo-fi models could involve only a subset of the relevant physical mechanisms, an approach popular, e.g. in climate science (Held, 2005; Majda & Gershgorin, 2010). Once specified, the models of different fidelities can be combined in different ways: one can either use the hi-fi model to periodically 'check' (and possibly adapt) the lo-fi models; or one can use the lo-fi models as pilot runs to decide whether or not to evaluate the hi-fi model; or one can use the information from all models simultaneously, by defining a multi-fidelity surrogate model (Craig et al., 1998; Cumming & Goldstein, 2009; Ehara & Guillas, 2023; Kennedy & O'Hagan, 2000), where correlation between models is taken into account. Provided that the lo-fi models are correlated with the original model, these additional cheap simulations can be leveraged to more accurately predict computer code output. The principal drawback of MFM is that there is limited guidance on how the lo-fi models should be constructed, and a poor choice can fail to improve (or even worsen) predictive performance, while incurring an additional computational cost. In addition, as with emulation, the literature on MFM tends to treat the hi-fi model as the target, rather than the continuum mathematical model.

Other Related Work. Some alternative lines of research will briefly be discussed. *Probabilistic numerics* casts numerical approximation as a statistical task (Hennig et al., 2015), with Bayesian principles used to quantify uncertainty regarding the continuum model of interest (Cockayne et al., 2019). However, the focus of the literature is the design of numerical methods, in contrast to extrapolation methods, which operate on the output of existing numerical methods. In parallel, the application of machine learning methods to numerical tasks has received recent

attention; for example, deep learning is being used for numerical approximation of high-dimensional parametric partial differential equations (Han et al., 2018). This literature does not attempt extrapolation as such, with a hi-fi numerical method typically used to provide a training dataset. Gaussian processes have been used in specific applications to extrapolate a series of numerical approximations to a continuum quantity of interest $f(0)$, for example, in Thodoroff et al. (2023) to model ice sheets in Antarctic, and in Ji et al. (2024) to model the evolution of the quark-gluon plasma following the Big Bang. To date, however, convergence acceleration has not been studied in the Gaussian process context. An important numerical task encountered in statistics is to approximate an expected value of interest $f(0) = \mathbb{E}[X(0)]$. Unbiased estimation of $f(0)$ at finite cost is possible in this setting using the methodology of Rhee and Glynn (2015), provided one can construct a sequence $(X(\mathbf{x}_n))_{n\in\mathbb{N}}$ of computable stochastic approximations to $X(0)$, such that the variance of $X(\mathbf{x}_n) - X(\mathbf{x}_{n-1})$ decays sufficiently fast. Similar de-biasing ideas have since been used in the context of Markov chain Monte Carlo (Jacob et al., 2020). Multilevel methods, based on such sequences, have been combined with Richardson extrapolation in Lemaire and Pagès (2017) and Beschle and Barth (2022).

Our Contribution. This paper proposes a probabilistic perspective on extrapolation methods that unifies extrapolation methods and multi-fidelity modelling (MFM). The approach is instantiated using a *numerical analysis-informed Gaussian process* to approximate the map $\mathbf{x} \mapsto f(\mathbf{x})$, as described in Section 2, where the conditional mean can be interpreted as a (novel) extrapolation method, in the sense that it provably achieves a polynomial (or even an exponential) speed-up compared to the original numerical method. Like Richardson extrapolation, our theoretical arguments are rooted in Taylor expansions, so the name Gauss–Richardson Extrapolation (GRE) is adopted. The probabilistic formulation of extrapolation methods confers several advantages:

- In contrast to classical extrapolation methods, which focus on the case of a univariate discretization parameter x_n, it is straight-forward to consider a vector of discretization parameters $\mathbf{x}_n$ within a regression framework. In Sections 2.1–2.3, the probabilistic approach is laid out, then in Sections 2.4 and 2.5, higher-order convergence guarantees for GRE are established.
- Credible sets for the continuum quantity of interest $f(0)$ can be constructed, enabling computational uncertainty to be integrated into experimental design and downstream decision-support. The asymptotic performance of GRE credible sets is analysed in Section 2.6.
- In contrast to existing approaches in MFM, where a discrete set of fidelities are specified at the outset, GRE operates on a continuous spectrum of fidelities and casts the selection of fidelities as a cost-constrained experimental design problem, which can then be approximately solved using methods described in Section 2.7.
- For computer models whose convergence order is difficult to analyse, the probabilistic formulation allows for convergence orders to be formally estimated. The consistency of a maximum quasi-likelihood approach to estimating unknown convergence order is established in Section 2.8.

The methodology is rigorously tested in the context of simulating from a computational cardiac model involving separate spatial and temporal discretization parameters in Section 3. The sensitivity of the cardiac model to the different discretization parameters is first estimated from lo-fi simulations, then an optimal experimental design is generated and used to estimate the true trajectory of the cardiac model in the continuum limit. Our experimental results demonstrate that a practical gain in accuracy can be achieved with our GRE method. Though our assessment focuses on a specific cardiac model of scientific and clinical interest, the methodology is general and offers an exciting possibility to accelerate computation in the diverse range of scenarios in which computationally intensive simulation is performed. A closing discussion is contained in Section 4.

Code to reproduce our results is provided at https://github.com/christopheroates/Richardson.

2 Methodology

This section contributes a probabilistic perspective on extrapolation methods, which we instantiate using Gaussian processes (GPs) to produce Gauss–Richardson Extrapolation (GRE). For

simplicity of presentation, we first consider the case of a scalar quantity of interest, generalizing to arbitrary-dimensional quantities of interest in Section 2.9.

Set-Up. Let $f : \mathcal{X} \to \mathbb{R}$ be a (nonrandom) real-valued function on a bounded set $\mathcal{X} \subset [0, \infty)^d$ such that $\mathbf{0} \in \mathcal{X}$. As in Section 1, the output $f(\mathbf{x})$ for $\mathbf{x} \neq \mathbf{0}$ will represent a numerical approximation to a continuum quantity $f(\mathbf{0})$ of interest, and for extrapolation to be possible at all we will minimally need to assume that f is continuous at $\mathbf{0}$.

Notation. For $\boldsymbol{\beta} \in \mathbb{N}_0^p$, let $\partial^{\boldsymbol{\beta}} g$ denote the mixed partial derivative $\mathbf{x} \mapsto \partial_{x_1}^{\beta_1} \dots \partial_{x_p}^{\beta_p} g(\mathbf{x})$ of a function $g : D \to \mathbb{R}$, whenever this is well-defined and $D \subseteq \mathbb{R}^p$. Let $C^s(D)$ denote the set of s-times continuously differentiable functions $g : D \to \mathbb{R}$, meaning that $\partial^{\boldsymbol{\beta}} g$ is continuous for all $|\boldsymbol{\beta}| \leq s$, where $|\boldsymbol{\beta}| := \beta_1 + \dots + \beta_p$. For $g : D \to \mathbb{R}$ bounded, let $\|g\|_{L^\infty(D)} := \sup_{\mathbf{x} \in D} |g(\mathbf{x})|$. Let $\pi_r(D)$ denote the set of all polynomial functions of total degree at most r on D. For vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, let $[\mathbf{a}, \mathbf{b}] := [a_1, b_1] \times \dots \times [a_d, b_d] \subset \mathbb{R}^d$, and similarly for $[\mathbf{a}, \mathbf{b})$ and so forth. Let $\mathcal{GP}(m, k)$ denote the law of a GP with mean function m and covariance function k; background on GPs can be found in Rasmussen and Williams (2006).

2.1 A numerical analysis-informed Gaussian process

Assuming for the moment that numerical analysis of $\mathbf{x} \mapsto f(\mathbf{x})$ can be performed, our first aim is to encode the resulting bounds on discretization error into a statistical regression model. Training such a *numerical analysis-informed* regression model on data $\{f(\mathbf{x}_i)\}_{i=1}^n$ obtained at distinct inputs $X_n = \{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X} \setminus \{\mathbf{0}\}$ enables statistical prediction of the limit $f(\mathbf{0})$, in analogy with classical extrapolation methods. To leverage conjugate computation, here we instantiate the idea using GPs in a Bayesian framework. For the first part, we require an explicit error bound $b : \mathcal{X} \to [0, \infty)$ such that $b(\mathbf{x}) \geq 0$ with equality if and only if $\mathbf{x} = \mathbf{0}$, and such that $f(\mathbf{x}) - f(\mathbf{0}) = O(b(\mathbf{x}))$. The error bound b will be encoded into a centred prior GP model for f, whose covariance function

$$k(\mathbf{x}, \mathbf{x}') := \sigma^2 \big[k_0^2 + b(\mathbf{x}) b(\mathbf{x}') k_e(\mathbf{x}, \mathbf{x}') \big], \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \tag{3}$$

is selected to ensure samples $g \sim \mathcal{GP}(0, k)$ from the GP satisfy, with probability one, $g(\mathbf{x}) - g(\mathbf{0}) = O(b(\mathbf{x}))$ (see Appendix B.1 for the precise statement). Here, $\sigma^2 > 0$ is an overall scale to be estimated, while the scalar $k_0^2 > 0$ is proportional to the prior variance of $f(\mathbf{0})$. The symmetric positive-definite function $k_e : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the covariance function for the *normalized error* $\mathbf{x} \mapsto e(\mathbf{x})$, where $e(\mathbf{x}) := b(\mathbf{x})^{-1}(f(\mathbf{x}) - f(\mathbf{0}))$ for $\mathbf{x} \in \mathcal{X} \setminus \{\mathbf{0}\}$, and must be specified. In practice, k_e will additionally involve length-scale parameters $\boldsymbol{\ell}$ which must be estimated, for example $k_e(\mathbf{x}, \mathbf{x}') = \exp(-\sum_{i=1}^d \ell_i^{-2}(x_i - x_i')^2)$ in the case of the Gaussian kernel; we defer all discussion of this point to Sections 2.8 and 3. To our knowledge, the encoding of convergence orders into a GP as in (3) has not been well-studied, though the basic idea appeared in Tuo et al. (2014) and Bect et al. (2021) and in our preliminary work (Teymur et al., 2021). Standard techniques can be applied to fit such a GP model to a dataset; see Figure 1 and Section 2.2.

Remark 1 (Recovering Richardson in dimension $d = 1$). Let k_e be any kernel that reproduces the polynomial space $\pi_{n-2}(\mathbb{R})$, such as $k_e(x, x') = (1 + xx')^{n-2}$, and consider the 'objective' prior with $k_0^2 \to \infty$. Conditioning on data $\{f(x_i)\}_{i=1}^n$, the posterior mean function is the unique interpolant of the form $x \mapsto \mu + b(x)p(x)$ for some $\mu \in \mathbb{R}$, $p \in \pi_{n-2}(\mathbb{R})$ (see e.g. Karvonen et al., 2018, Proposition 2.6). Thus, if b is polynomial, the intercept μ is the result of polynomial extrapolation to 0, and is an instance of Richardson's classical extrapolation method.

Unfortunately the connection in Remark 1 is not especially useful. Indeed, while the posterior mean provides a useful point estimate, the posterior variance is identically zero, meaning that predictive uncertainty is not being properly quantified. Thus, we do not attempt to reproduce Richardson extrapolation in the sequel, but rather, we develop *de novo* methodology tailored to the GP framework.

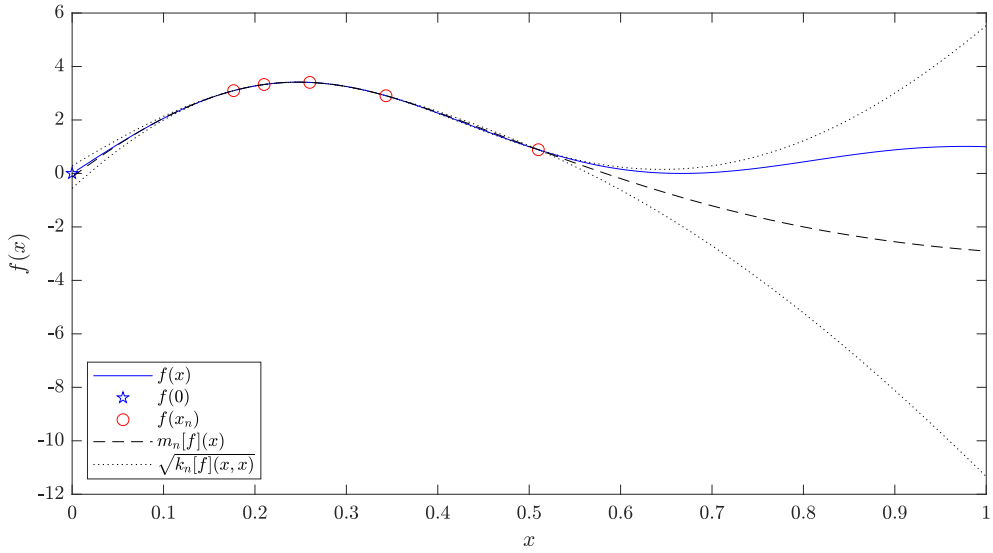

Figure 1. The numerical analysis-informed Gaussian process model, fitted to an illustrative dataset $\{f(x_i)\}_{i=1}^n$ (circles) of size $n = 5$, corresponding to the approximations produced by a finite difference method (solid curve) whose first-order accuracy [i.e. $b(x) = x$] was encoded into the GP. The scale $\sigma_n^2[f]$ of the uncertainty was calibrated using the method advocated in Section 2.6, while k_e was taken to be a Matérn-$\frac{5}{2}$ kernel with length-scale parameter selected using quasi-maximum likelihood (see Section 2.8). Observe that point estimate $m_n[f](0)$ (dashed curve at $x = 0$), is more accurate than that of the highest fidelity simulation from the numerical method, while the limiting quantity of interest $f(0)$ (star) falls within the one standard deviation prediction interval (dotted curves at $x = 0$).

2.2 Gauss–Richardson extrapolation

First we recall the relevant calculations for conditioning the GP model (3) on a dataset. Let $k_b : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be defined as $k_b(\mathbf{x}, \mathbf{x}') := b(\mathbf{x})b(\mathbf{x}')k_e(\mathbf{x}, \mathbf{x})$, so that our assumptions on b and k_e imply that k_b is a symmetric positive-definite kernel on $\mathcal{X} \setminus \{\mathbf{0}\}$, and a symmetric positive semi-definite kernel on $\mathcal{X}$. Let $f(X_n)$ be a column vector with entries $f(\mathbf{x}_i)$, let $\mathbf{k}_b(\mathbf{x})$ be a column vector with entries $k_b(\mathbf{x}_i, \mathbf{x})$, and let $\mathbf{K}_b$ be a matrix with entries $k_b(\mathbf{x}_i, \mathbf{x}_j)$. Recalling that k_0^2 is proportional to the prior variance for $f(\mathbf{0})$, we opt for an 'objective' prior in which $k_0^2 \to \infty$. However, this limit results in an improper prior GP. To make progress, we must first compute the conditional GP using a finite value of k_0^2 and then retrospectively take the limit—a standard calculation which we detail in Appendix B.2—yielding conditional mean and covariance functions

$$m_n[f](\mathbf{x}) := \frac{\mathbf{1}^\top \mathbf{K}_b^{-1} f(X_n)}{\mathbf{1}^\top \mathbf{K}_b^{-1} \mathbf{1}} + \mathbf{k}_b(\mathbf{x})^\top \mathbf{K}_b^{-1} \left\{ f(X_n) - \left(\frac{\mathbf{1}^\top \mathbf{K}_b^{-1} f(X_n)}{\mathbf{1}^\top \mathbf{K}_b^{-1} \mathbf{1}} \right) \mathbf{1} \right\}, \tag{4}$$

$$k_n[f](\mathbf{x}, \mathbf{x}') := \sigma_n^2[f] \left\{ k_b(\mathbf{x}, \mathbf{x}') - \mathbf{k}_b(\mathbf{x})^\top \mathbf{K}_b^{-1} \mathbf{k}_b(\mathbf{x}') + \frac{[\mathbf{k}_b(\mathbf{x})^\top \mathbf{K}_b^{-1} \mathbf{1} - 1][\mathbf{k}_b(\mathbf{x}')^\top \mathbf{K}_b^{-1} \mathbf{1} - 1]}{\mathbf{1}^\top \mathbf{K}_b^{-1} \mathbf{1}} \right\}, \tag{5}$$

where $\mathbf{1}$ is a column vector whose elements are all 1. The matrix $\mathbf{K}_b$ can indeed be inverted since we have assumed that the entries of $X_n \subset \mathcal{X} \setminus \{\mathbf{0}\}$ are distinct. To obtain (5), we have additionally replaced σ^2 with $\sigma_n^2[f]$, an estimator for the scale parameter σ, to be specified in Section 2.6. Computing the conditional mean and variance at $\mathbf{x} = \mathbf{0}$ results in the simple formulae

$$m_n[f](\mathbf{0}) = \frac{\mathbf{1}^\top \mathbf{K}_b^{-1} f(X_n)}{\mathbf{1}^\top \mathbf{K}_b^{-1} \mathbf{1}} \quad \text{and} \quad k_n[f](\mathbf{0}, \mathbf{0}) = \frac{\sigma_n^2[f]}{\mathbf{1}^\top \mathbf{K}_b^{-1} \mathbf{1}}, \tag{6}$$

since $b(\mathbf{0}) = 0$, and thus $k_b(\mathbf{0}, \mathbf{x}) = b(\mathbf{0})b(\mathbf{x})k_e(\mathbf{0}, \mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}$. The proposed GRE method returns a (univariate) Gaussian distribution, which can be summarized using the point estimate $m_n[f](\mathbf{0})$ for $f(\mathbf{0})$, together with the $100(1-\alpha)\%$ credible intervals

$$C_\alpha[f] = \left\{ y \in \mathbb{R} : \frac{|y - m_n[f](\mathbf{0})|}{\sqrt{k_n[f](\mathbf{0}, \mathbf{0})}} \leq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \right\}, \tag{7}$$

where Φ denotes the standard Gaussian cumulative density function. The uncertainty quantification provided by GRE unlocks additional functionality that was not available to classical extrapolation methods, including optimal experimental design for selecting X_n (Section 2.7) and principled statistical methods for estimating uncertain convergence orders (Section 2.8). However, both the accuracy of the point estimate and the coverage of the credible intervals will depend critically on the choice of the scale estimator $\sigma_n^2[f]$ and the choice of covariance function k_e. This important issue of how to select $\sigma_n^2[f]$ and k_e will be discussed next. An illustration of the proposed GRE method is provided in Figure 1.

2.3 Conservative Gaussian process priors

Our set-up involves a nonrandom function f that is modelled using a prior GP. One would perhaps hope to elicit a prior covariance function k in such a manner that f could plausibly have been generated as a sample from the GP. However, such elicitation is fundamentally difficult; the sample support set of a GP is not a vector space and may not even be measurable in general (Karvonen, 2023; Stein & Hung, 2019). How then can we proceed? In the applications that we have in mind, it is often possible to identify a symmetric positive-definite kernel such that f belongs to the reproducing kernel Hilbert space (RKHS) associated with the kernel, whose elements are real-valued functions on $\mathcal{X}$. For example, in numerical analysis, it is often possible to reason that f possesses a certain number of derivatives, from which inclusion in certain Sobolev RKHSs can be deduced. The approach that we take is to identify the covariance function k with the kernel of an RKHS, denoted $\mathcal{H}_k(\mathcal{X})$, in which f is contained. In particular, for any $k_0^2 \in (0, \infty)$ the space reproduced by the kernel k in (3) consists of functions $g : \mathcal{X} \to \mathbb{R}$ of the form $g(\mathbf{x}) = \mu + b(\mathbf{x})e(\mathbf{x})$ where $\mu \in \mathbb{R}$ and $e \in \mathcal{H}_{k_e}(\mathcal{X})$, and in the $k_0^2 \to \infty$ limit the norm structure of $\mathcal{H}_k(\mathcal{X})$ reduces to a semi-norm $|g|_{\mathcal{H}_k(\mathcal{X})} := \|\mathbf{x} \mapsto (g(\mathbf{x}) - g(\mathbf{0}))/b(\mathbf{0})\|_{\mathcal{H}_{k_e}(\mathcal{X})}$ induced by the norm structure of $\mathcal{H}_{k_e}(\mathcal{X})$; further background on RKHS can be found in Berlinet and Thomas-Agnan (2011). This construction results in a *conservative* prior GP, since with probability, one sample paths will be less regular than f when the RKHS is infinite-dimensional. However, there are several senses in which this approach to prior elicitation can be justified. First, it can be viewed as a form of 'objective' prior for GPs, in the sense that it is not intended to reflect prior belief but is rather intended to induce desirable behaviour in the posterior GP. Second, the choices that we make here will be justified through theoretical guarantees on both point estimation error (Section 2.4) and coverage of credible sets (Section 2.6). Third, the introduction of an additional scale estimator $\sigma_n^2[f]$ in (5) provides an opportunity to counteract the conservatism of the choice of k through the data-driven estimation of an appropriate scale for the credible sets in (7).

2.4 Higher-order convergence guarantees

The main technical contribution of this paper is to establish sufficient conditions under which the GRE point estimate $m_n[f](\mathbf{0})$ in (6) provides a more accurate approximation to the continuum limit $f(\mathbf{0})$ compared to the highest fidelity approximation $f(\mathbf{x}_n)$ on which it is based. The analysis we present is based on local polynomial reproduction, similar to that described in Wendland (2004). However, our results differ from existing work in that they are adapted to the nonstationary kernel (3) and quantify the space-filling properties of a design X_n using boxes, rather than balls or cones, since boxes are more natural for the domain $\mathcal{X} \subseteq [0, \infty)^d$ and enable sharper control over the constants involved.

To state our results, we define the *box fill distance* $\rho_{X_n,\mathcal{X}}$ as the supremum value of v such that there is a box of the form $[\mathbf{x}, \mathbf{x} + v\mathbf{1}]$ contained in $\mathcal{X}$ for which $X_n \cap [\mathbf{x}, \mathbf{x} + v\mathbf{1}] = \emptyset$. Define the constants γ_d using the induction $\gamma_d := 2d(1 + \gamma_{d-1})$ with base case $\gamma_1 := 2$. Our first main result, whose proof is contained in Appendix B.4, concerns the finite-smoothness case, where polynomial-order acceleration can be achieved:

Theorem 2 (Higher-order convergence; finite smoothness). Let $\mathcal{X} = [\mathbf{0}, \mathbf{1}] \subset \mathbb{R}^d$ and $X_n \subset \mathcal{X}$. Let $\mathcal{X}_h = [\mathbf{0}, h\mathbf{1}]$ and $X_n^h = \{h\mathbf{x} : \mathbf{x} \in X_n\}$ where $h \in (0, 1]$. Assume that $f \in \mathcal{H}_k(\mathcal{X})$, $b \in \pi_r(\mathcal{X})$ and $k_e \in C^{2s}(\mathcal{X} \times \mathcal{X})$. Let $m_n^h[f](\mathbf{0})$ denote the point estimate (6) based on data $f(X_n^h)$. Then there is an explicit n- and h-independent constant $C_{r,s}$, defined in the proof, such that

$$\underbrace{|f(\mathbf{0}) - m_n^h[f](\mathbf{0})|}_{\text{extrapolation error}} \leq C_{r,s} \rho_{X_n,\mathcal{X}}^s |f|_{\mathcal{H}_k(\mathcal{X})} \underbrace{h^s}_{\text{acceleration}} \underbrace{\|b\|_{L^\infty(\mathcal{X}_h)}}_{\text{original bound}}$$

whenever the box fill distance satisfies $\rho_{X_n,\mathcal{X}} \leq 1/(\gamma_d(r + 2s))$.

To interpret the conclusion of Theorem 2, fix n to be large enough that the constraint on the box fill distance is satisfied and examine the convergence of $m_n^h[f](\mathbf{0})$ to $f(\mathbf{0})$ as h is decreased. If the problem possesses no additional smoothness to exploit (i.e. $s = 0$) then convergence is gated at the rate $\|b\|_{L^\infty(\mathcal{X}_h)}$ of the original numerical method, irrespective of the number n of data that are used to train the GP. On the other hand, if f is regular enough that the normalized error functional $\mathbf{x} \mapsto (f(\mathbf{x}) - f(\mathbf{0}))/b(\mathbf{x})$ is an element of the RKHS $\mathcal{H}_{k_e}(\mathcal{X})$ of an s-smooth kernel (implied by $|f|_{\mathcal{H}_k(\mathcal{X})} < \infty$), then the h^s factor provides acceleration of polynomial order s over the convergence rate of the original numerical method. To the best of our knowledge, these theoretical results are the first of their kind for convergence acceleration using GPs. Examples 5 and 8 illustrate cases in which our regularity assumptions are satisfied. For the reader's convenience, we recall some standard examples of kernels and their associated smoothness properties in Appendix A.

Remark 3 (Sample efficiency compared to Richardson). A notable feature of Richardson extrapolation is that, under appropriate regularity assumptions, acceleration of order s can be achieved using a dataset of size $n = s + 1$ in dimension $d = 1$. For example, if f is first-order accurate with $f(h) = f(0) + c_1 h + O(h^2)$, then the line that passes through data $(h, f(h))$ and $(2h, f(2h))$ has intercept $2f(h) - f(2h)$, which is equal to $2[f(0) + c_1 h + O(h^2)] - [f(0) + 2c_1 h + O(h^2)] = f(0) + O(h^2)$; an additional order of accuracy is gained. Our result is less sample-efficient, in the sense that $n \geq 2r + 4s$ data are in principle required, due to the constraint on the box fill distance in Theorem 2. However, we speculate that this lower bound on n is not tight, and we empirically confirm that order-s acceleration is observed at smaller sample sizes n in Examples 5 and 8.

On the other hand, if there is infinite smoothness to exploit, then we may consider increasing the value of s in Theorem 2 to obtain an arbitrarily fast convergence rate as $h \to 0$, albeit with an increasing number n of training points required for the bound to hold. This result goes beyond classical Richardson extrapolation, but is natural within the GP framework. Theorem 4, whose proof is contained in Appendix B.5, is obtained by carefully tracking the s-dependent constants appearing in Theorem 2:

Theorem 4 (Higher-order convergence; infinite smoothness). In the setting of Theorem 2, assume further that $k_e \in C^\infty(\mathcal{X} \times \mathcal{X})$ and that $\sup_{\mathbf{x},\mathbf{y} \in \mathcal{X}} \sum_{|\boldsymbol{\beta}|=2s} |\partial_{\mathbf{y}}^{\boldsymbol{\beta}} k_e(\mathbf{x}, \mathbf{y})| \leq C_k^{2s}(2s)!$ for some constant C_k. Then, there exists an explicit h-independent constant $C_{n,r,s}$, defined in the proof, such that

$$\underbrace{|f(\mathbf{0}) - m_n^h[f](\mathbf{0})|}_{\text{extrapolation error}} \leq C_{n,r,s} |f|_{\mathcal{H}_k(\mathcal{X})} \underbrace{h^{\frac{1}{4\gamma_d \rho_{X_n,\mathcal{X}}}}}_{\text{acceleration}} \underbrace{\|b\|_{L^\infty(\mathcal{X}_h)}}_{\text{original bound}}$$

whenever the box fill distance satisfies $\rho_{X_n,\mathcal{X}} \leq \min\{1/(2\gamma_d(r + 1)), 1/(2d^{1/2}\gamma_d e^{4d\gamma_d + 1})\}$.

The derivative growth condition in the statement of Theorem 4 holds for most popular smooth kernels k_e, including the Gaussian kernel. The order of acceleration is now determined by the box

fill distance, which reflects the general phenomenon that 'more samples are required to exploit smoothness' (Cabannes & Vigogna, 2024).

To assess the sharpness of our results, we first consider the problem of approximating derivatives using finite differences; a setting where extrapolation methods are routinely used (see Section 6.7 of Brezinski & Zaglia, 2013):

Example 5 (Higher-order convergence for finite difference approximation). Consider numerical differentiation of a suitably regular function $\psi : \mathbb{R} \to \mathbb{R}$. The *central difference method*

$$f(x) := \frac{\psi(t+x) - \psi(t-x)}{2x}, \quad x > 0,$$

is a second-order approximation to $\psi'(t)$ for a given $t \in \mathbb{R}$. To make use of our results, we set $b(x) = x^2$, from (3), and suppose that $\psi(t+x) = c_0 + c_1 x + c_2 x^2 + c_3(x)x^3$ for some $c_0, c_1, c_2 \in \mathbb{R}$ and some x-dependent coefficient $c_3(x)$. The normalized error is

$$e(x) = \frac{f(x) - f(0)}{b(x)} = \frac{\psi(t+x) - \psi(t-x)}{2x \cdot x^2} - \frac{\psi'(t)}{x^2} = \frac{c_3(x) - c_3(-x)}{2},$$

so that the assumptions of Theorem 2 are satisfied when $x \mapsto c_3(x)$ and $x \mapsto c_3(-x)$ are elements of $\mathcal{H}_{k_e}(\mathcal{X})$, and $k_e \in C^{2s}(\mathcal{X} \times \mathcal{X})$ (the latter condition can be satisfied, for example, by taking k_e to be either a Matérn kernel or a Wendland kernel with appropriate smoothness level; see Appendix A). As a test problem, consider $\psi(t) = \sin(10t) + 1_{t>0} t^{s+4}$ with $\psi'(0) = 10$ the value to be estimated; this test problem is selected so that our assumptions hold precisely for an s-smooth kernel, as verified in Appendix B.6. The sample size $n = 5$ was fixed and the initial design $X_n = \{0.2, 0.4, 0.6, 0.8, 1\}$ was scaled by a factor h to obtain a range of designs $X_n^h \subset (0, h]$. In these experiments, we work in 100 digits of numerical precision, so that rounding error can be neglected.

Results for $s = 2$ are reported in Figure 2, with the *absolute error* $|f(0) - m_n^h[f](0)|$ plotted as a function of h in the left panel. These results reveal that the orders of acceleration predicted by our analysis are achieved, despite the sample size n being less than that required to fulfil the box fill distance requirement in Theorem 2. The GRE method demonstrated accuracy comparable to Richardson's extrapolation method (and superior to other classical extrapolation methods) when the kernel was chosen to match the smoothness of the task at hand. Interestingly, the most accurate extrapolation was provided by GRE with the Gaussian kernel, despite this kernel being too smooth for the task at hand. The coverage of GRE credible intervals was also investigated, with the *relative error* $(f(0) - m_n^h[f](0))/\sqrt{k_n[f](0,0)}$ plotted as a function of h in the right panel. It was found that credible intervals are asymptotically conservative in the case where a kernel with finite smoothness was used, in the sense that the relative error appeared to vanish in the $h \to 0$ limit. However, in the case of the Gaussian kernel, the credible intervals appeared to be asymptotically calibrated, in the sense that the relative error appeared to converge to a finite value (≈ 3) in the $h \to 0$ limit. Theoretical analysis of the GRE credible intervals is provided in Section 2.6.

Though they accurately describe the convergence acceleration provided by the GRE method, there are at least two apparent drawbacks with Theorems 2 and 4. The first is that these results require the error bound b to be a polynomial; this is an intrinsic part of our proof strategy, which is based on local polynomial reproduction, and cannot easily be relaxed. However, for applications in which a nonpolynomial error bound $\tilde{b}$ naturally arises, we may still be able to construct

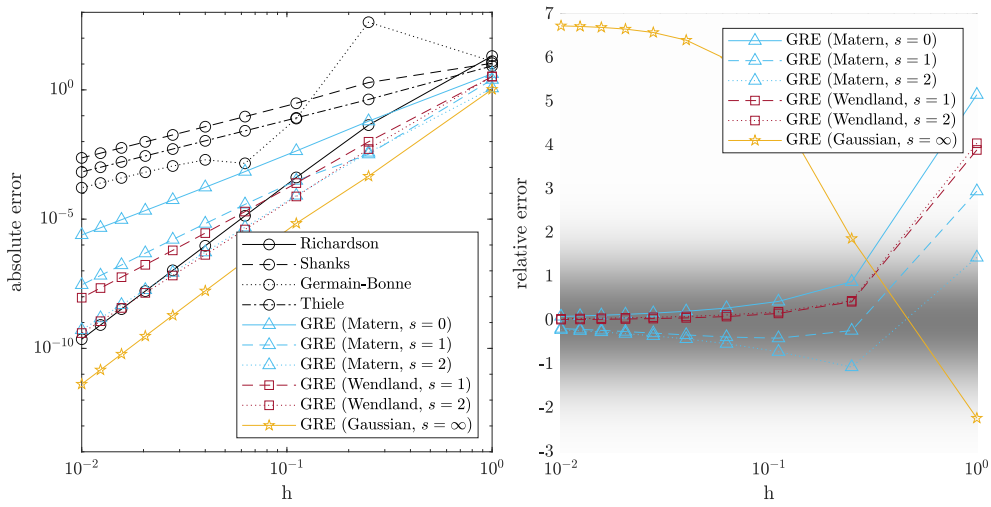

Figure 2. Accelerating the central difference method; Example 5. The left panel presents the absolute error $|f(0) - m_n^h[f](0)|$, while the right panel presents the relative error $(f(0) - m_n^h[f](0))/\sqrt{k_n[f](0,0)}$. Classical extrapolations methods (circles) were compared to our Gauss–Richardson Extrapolation (GRE) method, with either a Matérn (triangles), Wendland (squares), or Gaussian (stars) kernel. The true smoothness in this case is $s = 2$, while the legend indicates the level of smoothness assumed by the kernel. Kernel length-scale parameters were set to $\ell = 1$ and the scale estimator $\sigma_n^2[f]$ proposed in Section 2.6 was used. Shaded regions in the right panel correspond to the density function of the standard normal.

a polynomial error bound $b \in \pi_r(\mathcal{X})$ for some r that satisfies $\tilde{b}(\mathbf{x}) \leq b(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$ and enables the conclusion of Theorem 2 to be applied. The second limitation is that, for many iterative numerical methods that produce a convergent sequence of approximations to the continuum quantity $f(0)$ of interest, there is not always the notion of a continuum of discretization parameters $\mathbf{x}$ that can be exploited in the GRE framework. This second issue can be elegantly addressed using the notion of an s-smooth extension, which we introduce next.

2.5 The generality of continua

Several numerical methods do not admit a continuum of discretization parameters $\mathbf{x}$ that can be exploited in the GRE method. For example, the conjugate gradient algorithm for approximating the solution to a linear system of equations produces a convergent sequence of approximations, but is in no sense continuously indexed. The aim of this section is to demonstrate that iterative methods, which produce a sequence of approximations converging to a limiting quantity of interest, do in fact fall within our framework. The idea, roughly speaking, is to construct a function f whose values $f(x_n)$ on a convergent sequence, such as $x_n = 1/n$, coincide with the approximation produced after n iterations of the numerical method. The challenge is to show that such a function f exists with sufficient regularity that the results of Section 2.4 can be applied. Our main tool is the idea of an p-smooth extension, which is the content of Proposition 6. Let $\min(\mathbf{z}) := \min\{z_1, \ldots, z_d\}$ for $\mathbf{z} \in \mathbb{R}^d$.

Proposition 6 (p-smooth extension). Suppose that $C^p(\mathcal{X}) \subset \mathcal{H}_{k_e}(\mathcal{X})$ for some $p \in \mathbb{N}$. Let $(\mathbf{x}_n)_{n\in\mathbb{N}} \subset \mathcal{X} \setminus \{\mathbf{0}\}$ be such that $\mathbf{x}_{n+1} < \mathbf{x}_n$ componentwise and $\mathbf{x}_n \to \mathbf{0}$. Let $(y_n)_{n\in\mathbb{N}}$ be a convergent sequence with limit y_∞, such that the normalized errors $e_n := (y_n - y_\infty)/b(\mathbf{x}_n)$ satisfy $|e_n - e_{n+1}|/\min(\mathbf{x}_n - \mathbf{x}_{n+1})^p \to 0$. Then there exists a function f such that $f(0) = y_\infty$, $f(\mathbf{x}_n) = y_n$ for each $n \in \mathbb{N}$, and $|f|_{\mathcal{H}_k(\mathcal{X})} < \infty$.

A polynomial expansion can be used to establish the preconditions of Proposition 6, as we illustrate in the following result:

Corollary 7 (Sufficient conditions for p-smooth extension in $d = 1$). Let $(x_n, y_n)_{n\in\mathbb{N}} \subset (0, \infty) \times \mathbb{R}$ be such that x_n converges monotonically to 0, with $(x_n^{p+1} - x_{n+1}^{p+1})(x_n - x_{n+1})^{-p} \to 0$, $x_n^{p+2}(x_n - x_{n-1})^{-p} \to 0$ and $y_n = y_\infty + C_1 x_n^r + C_2 x_n^{r+p+1} + O(x_n^{r+p+2})$ for some constants $y_\infty, C_1, C_2 \in \mathbb{R}$. Let $b(x) = x^r$. Then, the preconditions of Proposition 6 are satisfied.

The proof of both Proposition 6 and Corollary 7 can be found in Appendix B.7. The conditions on the sequence $(x_n)_{n\in\mathbb{N}}$ in Corollary 7 are satisfied by, for example, sequences of the form $x_n = \frac{1}{n}$ and $x_n = \lambda^{-n}$ for any $\lambda > 1$, which are the sort of expressions that routinely appear in error bounds. The overall approach is illustrated in Example 8, where a GP analogue of the classical Romberg method for numerical integration is derived.

Example 8 (GP Romberg methods). Romberg methods for numerical integration are classically obtained via Richardson extrapolation of the trapezoidal rule (Brezinski & Zaglia, 2013, Section 6.7); it is interesting to ask if a similar feat can be achieved with GRE. Let $\psi \in C^{2m+2}([0, 1])$ and consider the trapezoidal rule $y_n := \frac{1}{n}[\frac{\psi(0)}{2} + \psi(\frac{1}{n}) + \ldots + \psi(\frac{n-1}{n}) + \frac{\psi(1)}{2}]$. The Euler–Maclaurin summation formula implies that the error of the trapezoidal rule can be expressed as

$$y_n - \int_0^1 \psi(t)\,\mathrm{d}t = \sum_{i=1}^{m} \frac{B_{2i}}{(2i)!} x_n^{2i}\left(\psi^{(2i-1)}(1) - \psi^{(2i-1)}(0)\right) + \frac{B_{2m+2}}{(2m+2)!} x_n^{2m+2} \psi^{(2m+2)}(\beta_n)$$

for some $\beta_n \in [0, 1]$, where $x_n = \frac{1}{n}$ and B_k are the Bernoulli numbers. As a test problem, consider $\psi(t) = \sin(10t) + t^2$, for which we can apply Corollary 7 with $b(x) = x^2$, $r = 2$ and $p = 3$. Thus there exists a function f that agrees with the trapezoidal rule on $(x_n)_{n\in\mathbb{N}}$ and satisfies the preconditions of Theorem 2 for a kernel k_e with smoothness up to $s = 2$; see Appendix A. Empirical results in Figure 3 verify that we are indeed able to gain an additional $s = 2$ convergence orders over the original trapezoidal rule, akin to Romberg integration, using our GRE method. Here, the sample size $n = 5$ was fixed and the initial design $X_n = \{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}\}$ was scaled by a factor h to obtain a range of designs $X_n^h \subset (0, h]$. The accuracy of the GRE point estimator and the coverage of the GRE credible interval demonstrate similar behaviour to that observed in Example 5.

These results extend the applicability of GRE to settings where components of the discretization parameter vector $\mathbf{x}$ could take values in any infinite set. For example, in standard implementations of the finite-element method for numerically solving partial differential equations one has a continuous parameter, characterizing the width of a triangular mesh, a discrete parameter, characterizing the number of cubature nodes used to integrate against each element, and another discrete parameter, specifying the number of iterations of a conjugate gradient method to solve the resulting linear system. The resulting mixture of continuous and discrete discretization parameters $\mathbf{x}$ falls within the scope of our GRE method.

2.6 Uncertainty quantification

An encouraging observation from Examples 5 and 8 was that the GRE credible intervals were not asymptotically over-confident as $h \to 0$. The aim of this section is to explain how the scale parameter σ^2 in (3), which controls the size of credible intervals $C_a[f]$ in (7), was actually estimated, and to rigorously prove that asymptotic over-confidence cannot occur when our proposed estimator $\sigma_n^2[f]$ is used.

The most standard approach to kernel parameter estimation is maximum (marginal) likelihood, but in GRE we do not have a valid likelihood due to taking the improper $k_0^2 \to \infty$ limit. Instead, we

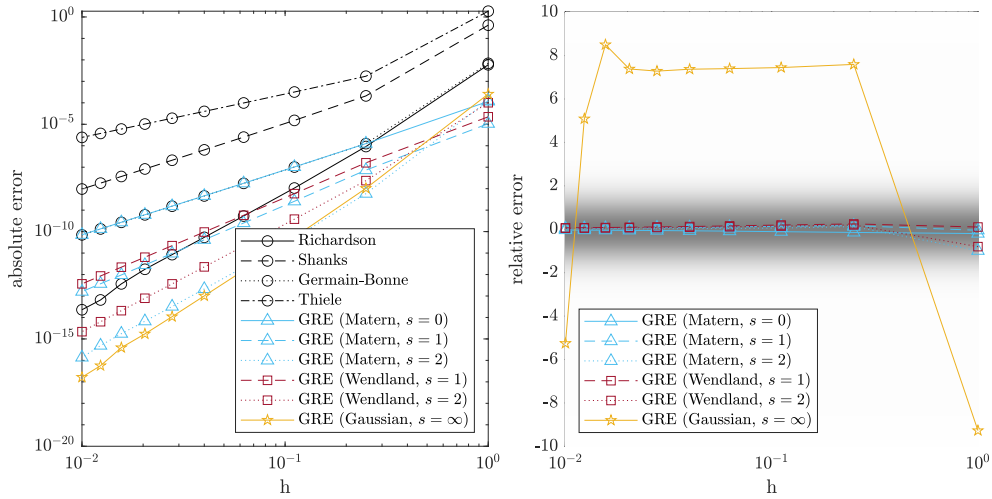

Figure 3. Accelerating the trapezoidal method to obtain a GP Romberg method; Example 8. The left panel presents the absolute error $|f(0) - m_n^h[f](0)|$, while the right panel presents the relative error $(f(0) - m_n^h[f](0))/\sqrt{k_n[f](0,0)}$. Classical extrapolations methods (circles) were compared to our GRE method, with either a Matérn (triangles), Wendland (squares), or Gaussian (stars) kernel. The true smoothness in this case is $s = 2$, while the legend indicates the level of smoothness assumed by the kernel. Kernel length-scale parameters were set to $\ell = 1$. Shaded regions in the right panel correspond to the density function of the standard normal.

motivate a particular estimator $\sigma_n^2[f]$ using asymptotic guarantees for the associated credible interval. Specifically, we advocate the estimator

$$\sigma_n^2[f] := \frac{|m_n[f]|^2_{\mathcal{H}_k(\mathcal{X})}}{n} = \frac{1}{n}\left[f(X_n)^\top \mathbf{K}_b^{-1} f(X_n) - \frac{(\mathbf{1}^\top \mathbf{K}_b^{-1} f(X_n))^2}{\mathbf{1}^\top \mathbf{K}_b^{-1}\mathbf{1}} \right], \tag{8}$$

which takes the same form as the maximum-likelihood estimator that we would have obtained had we not taken the $k_0^2 \to \infty$ limit, but with the semi-norm $|m_n[f]|_{\mathcal{H}_k(\mathcal{X})}$ in place of the conventional norm on $\mathcal{H}_k(\mathcal{X})$. This choice is supported by the following asymptotic result, whose proof is contained in Appendix B.8:

Proposition 9 (Asymptotic over-confidence is prevented). In the setting of Theorem 2, suppose that $s \geq 1$ and that $\lim_{\mathbf{x}\to\mathbf{0}} b(\mathbf{x})^{-1}(f(\mathbf{x}) - f(\mathbf{0})) \neq 0$ (i.e. we have a sharp error bound). Let $m_n^h[f](\mathbf{0})$ and $k_n^h[f](\mathbf{0},\mathbf{0})$ denote the conditional mean and variance in (6), based on data $f(X_n^h)$ and the estimator in (8). Then

$$\limsup_{h\to 0} \frac{|f(\mathbf{0}) - m_n^h[f](\mathbf{0})|}{\sqrt{k_n^h[f](\mathbf{0},\mathbf{0})}} < \infty$$

whenever the box fill distance $\rho_{X_n,\mathcal{X}}$ is sufficiently small.

In other words, the width $\sqrt{k_n[f](\mathbf{0},\mathbf{0})}$ of the credible interval cannot vanish asymptotically faster than the actual absolute error $|f(\mathbf{0}) - m_n[f](\mathbf{0})|$. Though this result does not guarantee that credible intervals are the 'right size' per se, there is no randomness in the data-generating process $f(\mathbf{x})$ and thus standard statistical notions of coverage, or 'right size', cannot be directly applied (see Karvonen et al., 2020). In practice, we have already seen empirical evidence that the credible sets (7) are appropriately conservative; an arguably predictable consequence of the conservative GP prior discussed in Section 2.3. Note that the conclusion of Proposition 9 also holds when the stronger hypotheses of Theorem 4 are assumed. However, the result assumes that a kernel

with appropriate smoothness is used; it does not explain the behaviour of GRE with the Gaussian kernel observed in Examples 5 and 8, since in that case the Gaussian kernel was formally misspecified.

Assured that our credible intervals are in a sense meaningfully related to the actual error, we can now proceed to exploit this measure of uncertainty for experimental design.

2.7 Optimal experimental design

One of the main engineering challenges associated with the simulation of continuum mathematical or physical phenomena is the numerical challenge of simultaneously controlling all sources of discretization error, to ensure the output $f(\mathbf{x})$ remains close in some sense to $f(\mathbf{0})$, the continuum quantity of interest. In practice, one might explore the sensitivity of the simulator output $f(\mathbf{x})$ to small changes in each discretization parameter x_i in turn, to heuristically identify a global setting $\mathbf{x}_{\text{hi-fi}}$ which is then fixed for the lifetime in which the simulator is used. It seems remarkable that more principled methodology has not yet been developed, and we aim to fill this gap by formulating *optimal experimental design* within the GRE framework.

The accuracy of the point estimator (6) will depend crucially on the locations at which the GP has been trained. Section 2.6 established that the conditional variance is meaningfully related to estimation accuracy, with the advantage that it can be explicitly calculated. This motivates the following cost-constrained optimization problem

$$\arg\max_{X\subset\mathcal{D}} \mathbf{1}^\top \mathbf{K}_b^{-1}\mathbf{1} \quad \text{s.t.} \quad \sum_{\mathbf{x}\in X} c(\mathbf{x}) \leq C, \tag{9}$$

where $\mathcal{D}\subseteq\mathcal{X}$ denotes the set of feasible simulations being considered, $\mathbf{K}_b$ is the matrix with entries $k_b(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{x}_i, \mathbf{x}_j \in X$, the map $c:\mathcal{D}\to\mathbb{R}$ quantifies the cost associated with obtaining simulator output $f(\mathbf{x})$, and C denotes the total computational budget. This numerical analysis-informed objective $\mathbf{1}^\top\mathbf{K}_b^{-1}\mathbf{1}$ is inversely proportional to the GRE posterior variance (6) when the scale parameter σ is fixed, rather than estimated (since a priori we do not suppose data have been obtained from which σ could be estimated). This optimization does not enforce a particular training sample size n, it just constrains the total computational cost. As such, (9) represents a challenging optimization problem, with both the number n of experiments in the optimal design, and the optimal experiments $X=\{\mathbf{x}_i\}_{i=1}^n$ themselves, to be determined. To proceed, we consider a finite set $\mathcal{D}$ of candidate experiments and then use brute force to search for an optimal design restricted to this candidate set, but we note that better search strategies can surely be developed.

Example 10 (Optimal experimental design in $d=1$). Consider a first-order numerical method with linear cost, so that $b(x)=x$ and $c(x)=x^{-1}$, an example of which would be the classical forward Euler method. For illustration, we take k_e to be either a Matérn kernel ($s=0$) or the Gaussian kernel ($s=\infty$), in each case with length-scale $\ell=1$ fixed. The total computational budget C was varied and optimal designs X were computed with elements constrained to a size 20 grid $\mathcal{D}$; results are shown in Figure 4. In the case of a rough kernel, like the Matérn kernel, a greedy/exploitative strategy of assigning all compute power to the highest resolution experiment seems optimal. Since we are working only with a discrete set of experiments, there is a small residual computational budget that is allocated to one or two further cheap experiments. For the Gaussian kernel, the optimal strategy is less greedy, with optimal designs involving more experiments, indicating that the greater smoothness is being leveraged to improve the accuracy of GRE.

In practice, a small number of preliminary simulations should be used to estimate appropriate length-scale parameters ℓ for the covariance kernel. Such parameter estimation becomes more critical in the multivariate setting, illustrated in the right panel of Figure 4, since the simulator output $f(\mathbf{x})$ may be more or less sensitive to different components of $\mathbf{x}$; in Section 3, a practical workflow is presented.

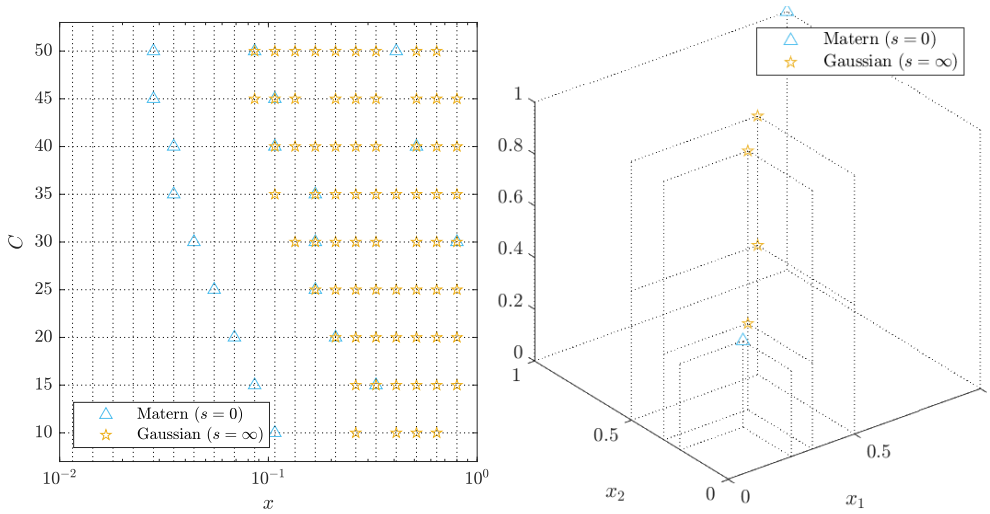

Figure 4. Optimal experimental designs were computed, for varying total computational budgets C, using either a Matérn (triangles; $s = 0$) or Gaussian (stars; $s = \infty$) kernel. Left: The setting of Example 10, with candidate states shown as vertical dotted lines on the plot. Right: An illustration of experimental design in dimension $d = 3$, with dotted lines used to indicates the coordinates of the states that were selected.

Remark 11 (Trivial solution for iterative methods). In Section 2.5, we discussed the scenario where data are generated along a sequence $(\mathbf{x}_n)_{n\in\mathbb{N}}$ by an iterative method, which first produces $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_{n-1})$ en route to producing the final output $f(\mathbf{x}_n)$. In this scenario, the cost of computing $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)$ is simply $c(\mathbf{x}_n)$, in which case computing as many iterations as possible is optimal in the sense of (9).

The methodology just presented systematizes the often *ad-hoc* process of selecting appropriate fidelities on which simulator output is computed, in a manner that is specifically tailored to improving the accuracy of our GRE method. Sequential experimental design strategies can also be developed, but were not pursued. The remainder of this section deals with three important generalizations of the GRE method; the case where convergence orders are unknown and must be estimated (Section 2.8), the case of multivariate simulator output (Section 2.9), and the case where the simulator contains additional degrees of freedom (Section 2.10).

2.8 Extension to unknown convergence order

The practical application of extrapolation methods does not necessarily require access to an explicit error bound, as several procedures have been developed to automatically identify a suitable method from a collection of extrapolation methods (which could correspond to different assumed convergence orders, or different classes of extrapolation method). A representative approach, called *automatic selection* (Delahaye, 1981), is based on the idea that small changes $S_{n+1} - S_n$ between consecutive iterates is a useful proxy for the convergence rate of an extrapolation method $(S_n)_{n\in\mathbb{N}}$. Another approach is to linearly combine estimates produced by a collection of extrapolation methods, called a *composite sequence approach* (Brezinski, 1985). From our statistical standpoint, these methods bear a respective semblance to model selection and model averaging. Pursuing a statistical perspective on extrapolation, here we consider maximum (marginal) likelihood as a default for selecting an appropriate GP prior model for GRE. The $k_0^2 \to \infty$ limit taken in Section 2.2 means that we do not have a proper likelihood, so instead, we identify and maximize an appropriate *quasi*-likelihood. Our justification is twofold, namely (1) our quasi-likelihood is directly analogous to the standard GP likelihood, and (2) we provide analysis below that demonstrates the consistency of maximum quasi-likelihood for estimation of convergence order in the GRE framework.

To formulate the main result of this section, we suppose we have a vector $\mathbf{r} \in \mathbb{R}^p$ that parametrizes the error bound $b_{\mathbf{r}} : \mathcal{X} \to [0, \infty)$, with the interpretation that increasing the value of any of the components of $\mathbf{r}$ corresponds to faster convergence of the error bound $b_{\mathbf{r}}(\mathbf{x})$ to 0 as $\mathbf{x} \to \mathbf{0}$. Specifically, we call a class of error bounds *monotonically parametrized* if, for all $\mathbf{r}_1 < \mathbf{r}_2$, we have

$$\inf_{\mathbf{r} \leq \mathbf{r}_1} \lim_{\mathbf{x} \to \mathbf{0}} \frac{b_{\mathbf{r}_2}(\mathbf{x})}{b_{\mathbf{r}}(\mathbf{x})} = 0.$$

This is not a restriction per se, as we are free to choose how $b_{\mathbf{r}}$ is parametrized, but an assumption of this kind is required to enable the following result to be rigorously stated. Examples of monotonically parametrized error bounds include $b_{\mathbf{r}}(\mathbf{x}) = x_1^{r_1} + \ldots + x_d^{r_d}$ and $b_{\mathbf{r}}(\mathbf{x}) = x_1^{r_1} \cdots x_d^{r_d}$, which are the sort of expressions that routinely appear in error bounds. The proof of the following result can be found in Appendix B.9:

Proposition 12 (Estimation using maximum quasi-likelihood). Let $X_n^h = \{h\mathbf{x} : \mathbf{x} \in X_n\}$. Suppose that $f \in \mathcal{H}_k(\mathcal{X})$ holds when k in (3) is based on the monotonically parametrized bound $b_{\mathbf{r}_0}(\mathbf{x})$ for some $\mathbf{r}_0 \geq \mathbf{0}$. Let $\mathbf{K}_{b_{\mathbf{r}},h}$ denote the matrix with entries $k_{b_{\mathbf{r}}}(h\mathbf{x}_i, h\mathbf{x}_j)$, where the dependence of this matrix on both h and $\mathbf{r}$ has now been emphasized, relative to the notation $\mathbf{K}_h$ introduced in Section 2.2. Then any maximizer $\mathbf{r}_n^h[f] \in \arg\max_{\mathbf{r} \geq 0} \mathcal{L}_n^h(\mathbf{r})$ of the log-quasi (marginal) likelihood

$$\mathcal{L}_n^h(\mathbf{r}) := -f(X_n^h)^{\top} \mathbf{K}_{b_{\mathbf{r}},h}^{-1} f(X_n^h) + \frac{(\mathbf{1}^{\top} \mathbf{K}_{b_{\mathbf{r}},h}^{-1} f(X_n^h))^2}{\mathbf{1}^{\top} \mathbf{K}_{b_{\mathbf{r}},h}^{-1} \mathbf{1}} - \log \det \mathbf{K}_{b_{\mathbf{r}},h} \tag{10}$$

satisfies $\liminf_{h \to 0} \mathbf{r}_n^h[f] \geq \mathbf{r}_0$.

The first two terms in (10) correspond to the (square of the) semi-norm $|m_n^h[f]|_{\mathcal{H}_k(\mathcal{X})}$, which is the analogue of the usual $\|m_n^h[h]\|_{\mathcal{H}_k(\mathcal{X})}$ term that would appear in the likelihood had we not taken the $k_0^2 \to \infty$ limit; this justifies the interpretation of (10), up to constants, as a quasi-likelihood. The one-sided conclusion of Proposition 12 may be surprising at first, but this is in fact the strongest result that can be expected. Indeed, the statement that $f(\mathbf{x}) - f(\mathbf{0}) = O(b_{\mathbf{r}_0}(\mathbf{x}))$ does not rule out the possibility that the error $f(\mathbf{x}) - f(\mathbf{0})$ decays *faster* than $b_{\mathbf{r}_0}(\mathbf{x})$, and in this case we would expect the estimator $\mathbf{r}_n^h[f]$ to adapt to the actual convergence order. The experiments that we report in Section 3 used maximum quasi (marginal) likelihood whenever convergence orders and/or kernel length-scale parameters were estimated.

Remark 13 (When to extrapolate?). The error bounds $b_{\mathbf{r}}(\mathbf{x})$ describe asymptotic behaviour as $\mathbf{x} \to \mathbf{0}$ only, and it is reasonable to ask whether given data $\{f(\mathbf{x}_i)\}_{i=1}^n$ are collected from a regime where such asymptotics are actually observed. Though we do not develop it further in this work, our statistical perspective enables goodness-of-fit testing and related techniques to assess the suitability of given data for being extrapolated.

2.9 Generalization to multidimensional output

Until this point, we have considered the continuum quantity of interest $f(\mathbf{0})$ to be scalar-valued. Oftentimes, however, we are interested in quantities $\{f(\mathbf{0}, \mathbf{t})\}_{\mathbf{t} \in \mathcal{T}}$ that are vector- or function-valued depending on the nature of the index set $\mathcal{T}$. The E-algorithm that we described in Section 1 has been extended to finite-dimensional vector-valued output; see Chapter 4 of Brezinski and Zaglia (2013) for detail. A possible advantage of the GP-based approach taken in GRE is that it does not impose any mathematical structure on $\mathcal{T}$ beyond this being a set, making extension of the methodology to function-valued output straight-forward.

To extend our methodology to multivariate output, let $f : \mathcal{X} \times \mathcal{T} \to \mathbb{R}$ be such that $\{f(\mathbf{0}, \mathbf{t})\}_{\mathbf{t} \in \mathcal{T}}$ is the continuum quantity of interest and $f(\mathbf{x}, \mathbf{t})$ is a numerical approximation to $f(\mathbf{0}, \mathbf{t})$. For

example, $f(0, t)$ may represent the solution to an ordinary differential equation at time t, while $f(x, t)$ may represent an approximation to this solution obtained using a Runge–Kutta method, with x being the error tolerance of the Runge–Kutta method. To improve presentation, we will assume that $f(\mathbf{x}, \mathbf{t}) - f(\mathbf{0}, \mathbf{t}) = O(b(\mathbf{x}))$ *uniformly* over $\mathbf{t} \in \mathcal{T}$, but $\mathbf{t}$-dependent error bounds could also be considered with additional notational overhead. Our original covariance function (3) can be generalized to

$$k((\mathbf{x}, \mathbf{t}), (\mathbf{x}', \mathbf{t}')) = \sigma^2\left[k_0^2 + b(\mathbf{x})b(\mathbf{x}')k_e(\mathbf{x}, \mathbf{x}')\right]k_{\mathcal{T}}(\mathbf{t}, \mathbf{t}'), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \mathbf{t}, \mathbf{t}' \in \mathcal{T}, \tag{11}$$

where to exploit tractable computation that results from this tensor product kernel, we have assumed a tensor product kernel and will assume that data $X_n = \{(\mathbf{x}_i, \mathbf{t}_j)\}_{i=1}^{n_1}{}_{j=1}^{n_2}$ are obtained on a Cartesian grid ($n = n_1 n_2$). That is, with the data appropriately ordered, we have the Kronecker decomposition $\mathbf{K} = \mathbf{K}_{\mathcal{X}} \otimes \mathbf{K}_{\mathcal{T}}$, where $\mathbf{K}_{\mathcal{X}}$ is the matrix with entries $k_0^2 + b(\mathbf{x}_i)b(\mathbf{x}_j)k_e(\mathbf{x}_i, \mathbf{x}_j)$, and $\mathbf{K}_{\mathcal{T}}$ is the matrix with entries $k_{\mathcal{T}}(\mathbf{t}_i, \mathbf{t}_j)$. Then, analogous calculations to those detailed in Appendix B.2, which we present in Appendix B.10, show that for values of $\mathbf{t}$, $\mathbf{t}'$ contained in the dataset, the conditional mean and covariance function in the $k_0 \to \infty$ limit are

$$\begin{aligned} m_n[f](\mathbf{x}, \mathbf{t}) &= \left\{\mathbf{k}_b(\mathbf{x})^\top \mathbf{K}_b^{-1} + [1 - \mathbf{k}_b(\mathbf{x})^\top \mathbf{K}_b^{-1}\mathbf{1}]\frac{\mathbf{1}^\top \mathbf{K}_b^{-1}}{\mathbf{1}^\top \mathbf{K}_b^{-1}\mathbf{1}}\right\} \otimes \left[k_{\mathcal{T}}(\mathbf{t})\mathbf{K}_{\mathcal{T}}^{-1}\right] f(X_n) \\ k_n[f]((\mathbf{x}, \mathbf{t}), (\mathbf{x}', \mathbf{t}')) &= \sigma_n^2[f]\left\{k_b(\mathbf{x}, \mathbf{x}')k_{\mathcal{T}}(\mathbf{t}, \mathbf{t}') - [\mathbf{k}_{\mathcal{T}}(\mathbf{t})^\top \mathbf{K}_{\mathcal{T}}^{-1}\mathbf{k}_{\mathcal{T}}(\mathbf{t}')]\right. \\ &\quad \left.\times \left[\mathbf{k}_b(\mathbf{x})^\top \mathbf{K}_b^{-1}\mathbf{k}_b(\mathbf{x}') - \frac{[\mathbf{k}_b(\mathbf{x})^\top \mathbf{K}_b^{-1}\mathbf{1} - 1][\mathbf{k}_b(\mathbf{x}')^\top \mathbf{K}_b^{-1}\mathbf{1} - 1]^\top}{\mathbf{1}^\top \mathbf{K}_b^{-1}\mathbf{1}}\right]\right\}, \end{aligned}$$

where $\mathbf{k}_{\mathcal{T}}(\mathbf{t})$ is the vector with entries $k_{\mathcal{T}}(\mathbf{t}_i, \mathbf{t})$. For values of $\mathbf{t}$, $\mathbf{t}'$ not contained in the training dataset, the conditional covariance does not have a finite limit; a proper prior should be used if off-grid prediction in the $\mathbf{t}$-domain is required. Further details on the multivariate setting are deferred to Section 3, where the approach is explored in the context of predicting temporal output from a cardiac model.

2.10 Incorporating additional degrees of freedom

The final methodological extension that we consider is the case where $f_{\boldsymbol{\theta}}(\mathbf{x})$ additionally depends on one or more degrees of freedom $\boldsymbol{\theta} \in \Theta$; a setting where emulation or MFM methods are routinely used (cf. Section 1). The proposed GRE method can be applied in this context by viewing $f_{\boldsymbol{\theta}}(\mathbf{0})$ as a simulator with multidimensional output $\{f(\mathbf{0}, \boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$ and then applying the methodology described in Section 2.9 with $\boldsymbol{\theta}$, rather than $\mathbf{t}$, indexing the output of this extended model. Since the required calculations are identical, we do not dwell any further on this point.

This completes our exposition of the GRE method. Next, we next turn to a cardiac modelling case study, where the usefulness of the methodology is evaluated.

3 Case study: cardiac modelling

The cardiac model $f_{\boldsymbol{\theta}}(\mathbf{x})$ that we consider in this section is a detailed numerical simulation of a single heart beat[1] (Strocchi et al., 2023). The simulation is rooted in finite element methods that require both a spatial (x_1) and a temporal (x_2) discretization level to be specified; of these, the spatial discretization is the most critical, due to the $O(x_1^{-3})$ cost associated with the construction of a suitable triangulation of the time-varying 3-dimensional volume of the heart; see Figure 5. The computational cost $c(\mathbf{x})$ is measured in real computational time (seconds) and comprises the *setup time*, *assembly time* (the time taken to assemble linear systems of equations), and the *solver time* (the time taken to solve linear systems of equations), with assembly time the main contributor

[1] The simulation is usually run until a steady state is reached before reading off quantities of interest, at a substantial increase to the overall computational cost. For the present purpose, we removed components from the model that required multiple heart beats to reach a steady state, and simulated only a single heart beat.

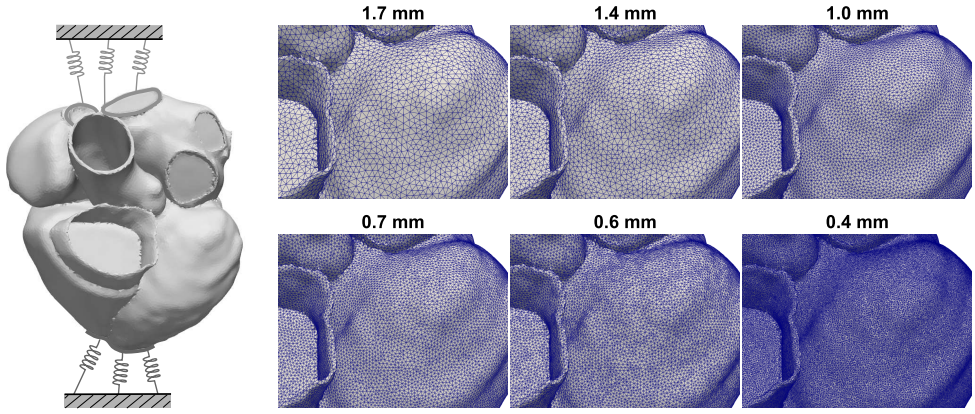

Figure 5. Cardiac model: Left: Schematic indicating the veins and the apical region where spring boundary conditions were applied. Right: A subset of the mesh resolutions used in this case study. The finest resolution required 3×10^7 finite elements to be used.

to total computational cost. To achieve a clinically acceptable level of accuracy, it is typical for a simulation $f_{\boldsymbol{\theta}}(\mathbf{x}_{\text{default}})$ to be performed with $\mathbf{x}_{\text{default}} \approx (0.4\,\text{mm}, 2\,\text{ms})$, at a cost $c(\mathbf{x}_{\text{default}}) \approx 1.5 \times 10^4$ s (around ≈ 4 hr) for a single heart beat.[2] This poses severe challenges to the scientific use of such models, with super-computing resources required to ascertain whether there are values of scientific parameters $\boldsymbol{\theta}$ for which observed data are consistent with model output (Strocchi et al., 2023). These challenges directly motivated the development of GRE, and the remainder of this paper is dedicated to exploring the value of extrapolation methods in this context. Extrapolation of the cardiac model output represents a much greater challenge compared to extrapolation for the examples considered in Section 2, due to the nonlinear physics being simulated. Since our focus in this paper is not on inference for $\boldsymbol{\theta}$, these degrees of freedom were fixed to physically realistic values based on previous analyses (Strocchi et al., 2020, 2023), with all further details on the construction of the cardiac model reserved for Appendix C.

Section 3.1 sets out a practical workflow for using the GRE method, that focuses on the multi-dimensional setting where both convergence orders and kernel length-scale parameters are to be estimated. The performance of GRE is then investigated for both scalar-valued (Section 3.2) and multivariate (Section 3.3) continuum quantities of interest.

3.1 A proposed general workflow

The sophistication of the cardiac model renders analytical derivation of convergence orders essentially impossible, so to proceed these orders must be estimated. However, the computational cost of simulating from the model means that data from which convergence orders can be estimated are necessarily limited. This motivates us to propose the following pragmatic workflow, which we present for a general model $f(\mathbf{x})$ and which scales in a reasonable way with the number d of components of $\mathbf{x}$ that can be varied. This workflow requires the user to specify a lo-fi setting $\mathbf{x}_{\text{lo-fi}}$ as a starting point, together with a means to predict the computational cost $c(\mathbf{x})$ of simulating $f(\mathbf{x})$, and a total computational budget C:

1. For each fidelity parameter x_i, $i = 1, \ldots, d$:
   (a) Simulate $f(\mathbf{x})$ for a range of values of x_i, with all of the other components $\mathbf{x}$ held fixed to their values in $\mathbf{x}_{\text{lo-fi}}$.
   (b) Fit a univariate numerical analysis-informed GP model (4), (5) to these data, assuming an error bound of the form $b(x_i) = x_i^{r_i}$, where the scale estimate $\hat{\sigma}_i$ from Section 2.6 is used,

[2] Simulations for this case study were performed on ARCHER2, a UK national super computing service (https://www.archer2.ac.uk/). Each simulation involved 512 CPUs operating in parallel, so that simulation of one heart beat using setting $\mathbf{x}_{\text{default}}$ required $\approx 4 \times 512$ CPU hours in total.

and where the convergence order r_i, the kernel smoothness s_i, and the kernel length-scale parameter ℓ_i are simultaneously estimated using quasi maximum likelihood, as explained in Section 2.8.

2. Construct a tensor product covariance model $k_e(\mathbf{x}, \mathbf{x}') = k_e(x_1, x_1'; \ell_1) \dots k_e(x_d, x_d'; \ell_d)$ and posit the overall error bound $b(\mathbf{x}) = \hat{\sigma}_1 x_1^{r_1} + \dots + \hat{\sigma}_d x_d^{r_d}$. Then, perform experimental design as described in Section 2.7, with computational budget C. Denote the optimal design X_n.
3. Simulate $f(\mathbf{x})$ for each $\mathbf{x} \in X_n$ and return the GRE conditional mean (6) as the final approximation to $f(\mathbf{0})$.

Several remarks are in order: First, it is assumed that the Step 1 incurs negligible cost relative to the total computational budget; the precise interpretation of this assumption will necessarily be context-dependent. Second, the additive form for $b(\mathbf{x})$ is appropriately conservative, in the sense that *all* components of $\mathbf{x}$ must be small to control this bound. One could go further and compare the performance of GPs based on alternative form of $b(\mathbf{x})$, for example, with interaction terms included, selecting among such models using maximum quasi-likelihood, but for the present purposes the additive form of $b(\mathbf{x})$ is preferred since it is compatible with the independent estimation of convergence orders r_i in Step 1. Third, the independent estimation of (r_i, s_i, ℓ_i) for each $i = 1, \dots, d$ can be performed using brute-force search over a 3-dimensional grid to maximize the quasi-likelihood, whereas simultaneous estimation of all kernel parameters would be both statistically and computationally difficult. The full workflow is demonstrated on our cardiac case study, next.

3.2 Approximation of scalar quantities of interest

The first part of our case study concerned the approximation of physiologically interpretable scalar-valued quantities of interest. These were the minimum volume of the left and right ventricles and atria, the maximum volume during ventricular contraction for the left and right atria, and the time taken for the ventricles to contract in total capacity by one-half; a total of seven test problems for GRE.

Though the computational time $c(\mathbf{x}_{\text{default}})$ is substantial, in this case, study parallel computation resources can be exploited. The main computational constraint that we work under here is that we will only run experiments for which $c(\mathbf{x}) \leq c(\mathbf{x}_{\text{default}})$ within our GRE method. To circumvent the complication of predicting computational times before experiments are performed, for this case study, a discrete set of experiments were performed at the outset and their times recorded. Since the continuum limit $f(\mathbf{0})$ is intractable, we additionally computed a reference solution $f(\mathbf{x}_{\text{hi-fi}})$ with $\mathbf{x}_{\text{hi-fi}} = (0.4\text{ mm}, 1\text{ ms})$ and in what follows we assess how well the GRE point estimate $m_n[f](\mathbf{x}_{\text{hi-fi}})$ approximates $f(\mathbf{x}_{\text{hi-fi}})$. The central question here is whether the workflow proposed in Section 3.1 can provide more accurate approximation of $f(\mathbf{x}_{\text{hi-fi}})$ compared to $f(\mathbf{x}_{\text{default}})$, and if so what computational budget is required. To the best of our knowledge, there do not exist comparable methodologies for this task; methods such as emulation and MFM are not applicable when $\boldsymbol{\theta}$ is fixed, and classical extrapolation methods were not developed with multivariate $\mathbf{x}$ in mind.

The workflow is illustrated in the left panel of Figure 6. The lo-fi setting was $\mathbf{x}_{\text{lo-fi}} = (1.7\text{ mm}, 5\text{ ms})$. The convergence orders r_1, r_2 were selected from $\{0.5, 1, 2\}$, the smoothness parameters s_1, s_2 were selected from $\{0, 1, 2\}$, and the length-scales ℓ_1, ℓ_2 were selected using grid search, all estimated simultaneously using maximum quasi-likelihood. Experimental designs were computed based on a candidate set of experiments, each of which incurs a cost no greater than $c(\mathbf{x}_{\text{default}})$, indicated by dots in the left panel of Figure 6. Results for the seven test problems are shown in the right panel of Figure 6, where it is observed that the GRE point estimator provides a generally better approximation to $f(\mathbf{x}_{\text{hi-fi}})$ compared to $f(\mathbf{x}_{\text{default}})$ when the computational budget C reaches or exceeds 10^5. The optimal design for approximating the minimum volume of the left ventricle is depicted in the left-hand panel of Figure 6 for a computational budget $C = 10^5$; the design supplements $\mathbf{x}_{\text{default}}$ with 6 additional simulations of lower cost, analogous to a classical extrapolation method but here generalized to the multivariate context. Note that for C exceeding 2×10^5 the optimal design becomes saturated, containing all experiments in the candidate set. That GRE should perform worse than $f(\mathbf{x}_{\text{defult}})$ at small computational budgets is not surprising given that all convergence orders r_i, smoothnesses s_i, and length-scales ℓ_i are estimated from the

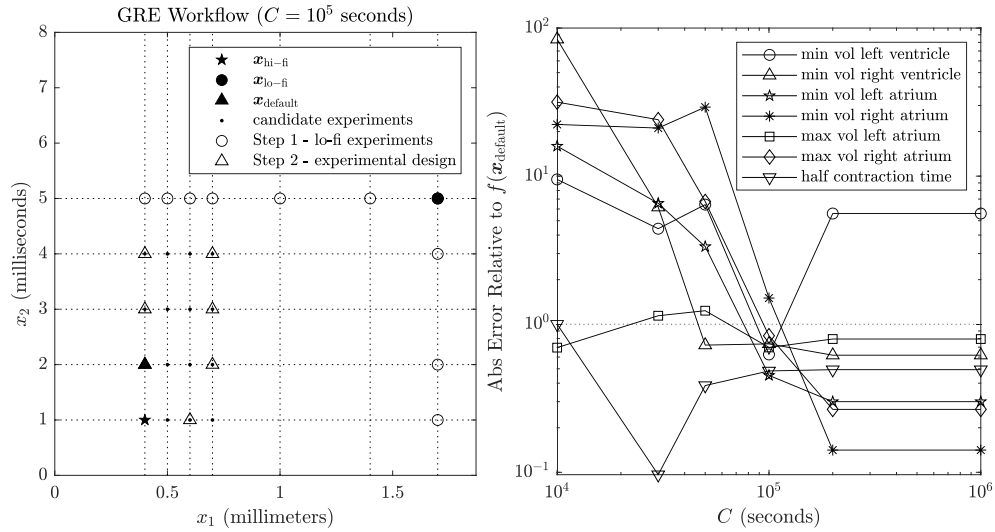

Figure 6. Scalar quantities of interest from the cardiac model. Left: The workflow, illustrated. In Step 1, the effect of varying each component of $\mathbf{x}$ in turn is explored, with all other components fixed equal to their value in $\mathbf{x}_{\text{lo-fi}}$. This facilitates the construction of a multivariate Gaussian process model for use in Step 2, where experimental design is performed (here shown for a computational budget of $C = 10^5$ s). For assessment purposes we aim to predict $f(\mathbf{x}_{\text{hi-fi}})$ as a ground truth, but in practice the goal is to predict $f(\mathbf{0})$. Right: For each of the seven scalar quantities of interest associated with the cardiac model we display the ratio of the absolute error $|f(\mathbf{x}_{\text{hi-fi}}) - m_n[f](\mathbf{x}_{\text{hi-fi}})|$ of the GRE method and the absolute error $|f(\mathbf{x}_{\text{hi-fi}}) - f(\mathbf{x}_{\text{default}})|$ of the default approximation, as a function of the total computational budget C.

small lo-fi training dataset, and these values largely determine the output of GRE in the absence of a sufficient number of experiments in the training dataset X_n. However, for a sufficiently large computational budget, it is encouraging to see that information from the experiments in X_n, each of which cost no greater than $c(\mathbf{x}_{\text{default}})$, is exploited in GRE to achieve more accurate estimation for 6 of the 7 scalar quantities of interest.

3.3 Approximation of temporal model output

The scalar quantities of interest considered in Section 3.2 are summary statistics obtained from 4-dimensional temporal model output of the form $\mathbf{f}(\mathbf{x}, t)$, where here t is a time index ranging from 0 to 600 ms and the components of $\mathbf{f}$ refer to the volumes of the atria and ventricles. It is, therefore, interesting to investigate whether these temporal outputs can be directly approximated, providing four test problems for the methodology described in Section 2.9. Here, for simplicity, we fixed the discrete values s_1, s_2, r_1, and r_2, the median of the values estimated in Section 3.2, and we fixed the continuous values $\hat{\sigma}_1$, $\hat{\sigma}_2$, ℓ_1, and ℓ_2 to the mean of the values estimated in Section 3.2. The length-scale for the kernel $k_{\mathcal{T}}$ was set equal to the length of the time series itself. The computational budget was fixed to $C = 2 \times 10^5$, so that our experimental design is saturated, but recall that no individual experiment in this design had cost exceeding $c(\mathbf{x}_{\text{default}})$. Full results are displayed in Figure 7. In each case, the approximation produced by GRE achieves lower mean square error relative to $f(\mathbf{x}_{\text{default}}, t)$. Taken together with the results in Section 3.2, these results are an encouraging and pave the way for subsequent investigations and applications of GRE.

4 Discussion

This paper introduced a probabilistic perspective on extrapolation, presenting a framework in which classical extrapolation methods from numerical analysis and modern MFM are unified. One approach was developed in detail, which we termed GRE. The GRE method facilitates simultaneous convergence acceleration and uncertainty quantification, and unlocks experimental design functionality for optimization over the set of fidelities at which simulation is performed. The end result is a methodology that allows a practitioner to arrive, in a principled manner, at fidelities

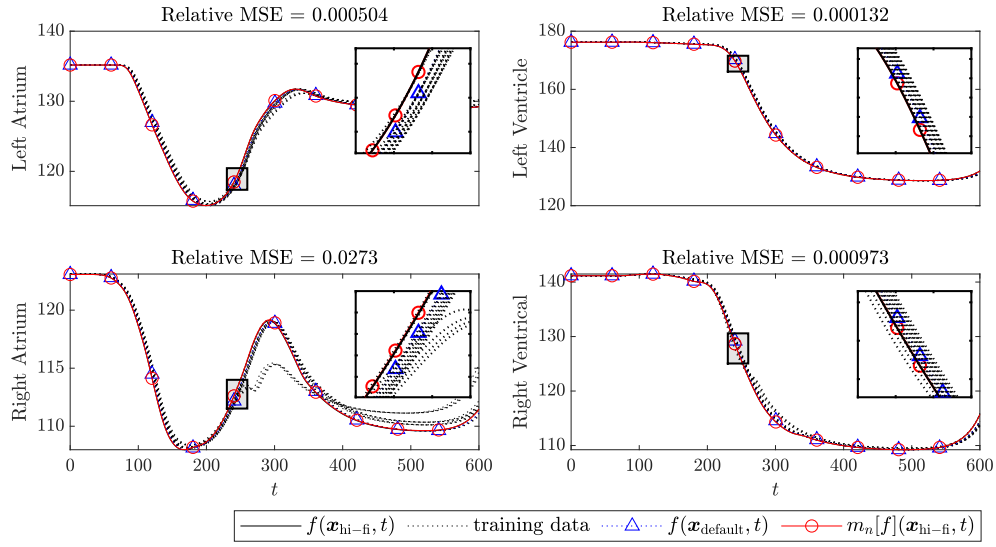

Figure 7. Temporal quantities of interest from the cardiac model. For each of the four temporal quantities of interest associated with the cardiac model we display the approximations produced at the different spatial and temporal resolutions in the training dataset X_n, together with the ground truth $f(\mathbf{x}_{\text{hi-fi}})$ (solid), the default approximation $f(\mathbf{x}_{\text{default}}, t)$ (triangles), and the approximation $m_n[f](\mathbf{x}_{\text{hi-fi}}, t)$ from Gauss–Richardson extrapolation (circles). The ratio of the mean square error $\int [f(\mathbf{x}_{\text{hi-fi}}, t) - m_n[f](\mathbf{x}_{\text{hi-fi}}, t)]^2 \, \mathrm{d}t$ of the GRE method and the mean square error $\int [f(\mathbf{x}_{\text{hi-fi}}, t) - f(\mathbf{x}_{\text{default}}, t)]^2 \, \mathrm{d}t$ of the default method is also reported.

$\{\mathbf{x}_i\}_{i=1}^n$ such that the associated simulator outputs $\{f(\mathbf{x}_i)\}_{i=1}^n$ can be combined to produce an approximation to the continuum quantity $f(\mathbf{0})$ that is typically more accurate than a single hi-fi simulation run at a comparable computational cost. A cardiac modelling case study provided an initial positive proof-of-concept, but further case studies—involving different types of computer model—will be required to comprehensively assess GRE; we aim to undertake domain-specific investigations in future work.

Several methodological extensions to this work can be envisaged, such as considering alternative regression models to GPs, developing theory and methodology for the more challenging cases where the regression model is misspecified and computational costs needs to be predicted, and extending the experimental design methodology to include additional degrees of freedom $\boldsymbol{\theta}$, which are often present in a mathematical model. In addition, and more speculatively, it would be interesting to explore modern computational tasks, such as the super-resolution task in deep learning, for which extrapolation methods have yet to be exploited.

Acknowledgments

The authors wish to thank Julien Bect, Simon Mak, Onur Teymur, and Jere Koskela for discussions of this project.

Conflicts of interest: None declared.

Funding

C.J.O. was supported by EP/W019590/1 and a Leverhulme Prize. T.K. was supported by the Research Council of Finland postdoctoral researcher grant number 338567, *Scalable, Adaptive and Reliable Probabilistic Integration*. A.L.T. was supported by EP/X01259X/1 and EP/R014604/1. C.J.O. and A.L.T. would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme Mathematical and Statistical Foundation of Future Data-Driven Engineering where work on this paper was undertaken. C.J.O. and S.A.N. were supported by The Alan Turing Institute, UK. M.S. and S.A.N.

were supported by the Wellcome/EPSRC Centre for Medical Engineering (WT203148/Z/16/Z). S.A.N. was supported by NIH R01-HL152256, ERC PREDICT-HF 453 (864055), BHF (RG/20/4/34803), and EPSRC (EP/P01268X/1, EP/X03870X/1).

Data availability

The data that support the findings of this study are available on GitHub.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Arroyo C. P., Dombard J., Duchaine F., Gicquel L., Martin B., Odier N., & Staffelbach G. (2021). Towards the large-eddy simulation of a full engine: Integration of a 360 azimuthal degrees fan, compressor and combustion chamber. Part I: Methodology and initialisation. *Journal of the Global Power and Propulsion Society*, 133115. https://doi.org/10.33737/jgpps/133115

Bach F. (2021). On the effectiveness of Richardson extrapolation in data science. *SIAM Journal on Mathematics of Data Science*, *3*(4), 1251–1277. https://doi.org/10.1137/21M1397349

Bect J., Zio S., Perrin G., Cannamela C., & Vazquez E. (2021). On the quantification of discretization uncertainty: Comparison of two paradigms. In *14th World Congress in Computational Mechanics and ECCOMAS Congress 2020 (WCCM-ECCOMAS)*.

Berlinet A., & Thomas-Agnan C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Beschle C. A., & Barth A. (2022). Quasi continuous level Monte Carlo for random elliptic PDEs. In *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 3–31. Springer.

Brezinski C. (1985). Composite sequence transformations. *Numerische Mathematik*, *46*(3),311–321. https://doi.org/10.1007/BF01389488

Brezinski C. (1989). A survey of iterative extrapolation by the E-algorithm, Det Kong. *Det Kongelige Norske Videnskabers Selskabs Skrifter*, *2*, 1–26.

Brezinski C., & Zaglia M. R. (2013). *Extrapolation methods: Theory and practice*. Elsevier.

Bulirsch R., & Stoer J. (1964). Fehlerabschätzungen und extrapolation mit rationalen funktionen bei verfahren vom Richardson-typus. *Numerische Mathematik*, *6*(1), 413–427. https://doi.org/10.1007/BF01386092

Cabannes V., & Vigogna S. (2024). How many samples are needed to leverage smoothness? In *Proceedings of the 38th Conference on Neural Information Processing Systems*.

Chizat L., Roussillon P., Léger F., Vialard F.-X., & Peyré G. (2020). Faster Wasserstein distance estimation with the Sinkhorn divergence. In *Proceedings of the 34th Conference on Neural Information Processing Systems*.

Cockayne J., Oates C. J., Sullivan T. J., & Girolami M. (2019). Bayesian probabilistic numerical methods. *SIAM Review: Society for Industrial and Applied Mathematics*, *61*(3), 756–789. https://doi.org/10.1137/17M1139357

Craig P. S., Goldstein M., Seheult A., & Smith J. (1998). Constructing partial prior specifications for models of complex physical systems. *Journal of the Royal Statistical Society, Series D*, *47*(1), 37–53. https://doi.org/10.1111/1467-9884.00115

Cumming J. A., & Goldstein M. (2009). Small sample Bayesian designs for complex high-dimensional models based on information gained using fast approximations. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, *51*(4), 377–388. https://doi.org/10.1198/TECH.2009.08015

Delahaye J.-P. (1981). Automatic selection of sequence transformations. *Mathematics of Computation*, *37*(155), 197–204. https://doi.org/10.1090/mcom/1981-37-155

Durmus A., Simsekli U., Moulines E., Badeau R., & Richard G. (2016). Stochastic gradient Richardson-Romberg Markov chain Monte Carlo. In *Proceedings of the 30th Conference on Neural Information Processing Systems*.

Ehara A., & Guillas S. (2023). An adaptive strategy for sequential designs of multilevel computer experiments. *International Journal for Uncertainty Quantification*, *13*(4),61–98. https://doi.org/10.1615/Int.J.UncertaintyQuantification.v13.i4

Germain-Bonne B. (1990). Convergence acceleration of number-machine sequences. *Journal of Computational and Applied Mathematics*, *32*(1-2), 83–88. https://doi.org/10.1016/0377-0427(90)90419-Z

Han J., Jentzen A., & E W. (2018). Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, *115*(34), 8505–8510. https://doi.org/10.1073/pnas.1718942115

Held I. M. (2005). The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society*, *86*(11), 1609–1614. https://doi.org/10.1175/BAMS-86-11-1609

Hennig P., Osborne M. A., & Girolami M. (2015). Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *471*(2179), 20150142. https://doi.org/10.1098/rspa.2015.0142

Jacob P. E., O'Leary J., & Atchadé Y. F. (2020). Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society, Series B*, *82*(3), 543–600. https://doi.org/10.1111/rssb.12336

Ji Y., Yuchi H. S., Soeder D., Paquet J.-F., Bass S. A., Joseph V. R., Wu C. J., & Mak S. (2024). Conglomerate multi-fidelity Gaussian process modeling, with application to heavy-ion collisions. *SIAM/ASA Journal on Uncertainty Quantification*, *12*(2), 473–502. https://doi.org/10.1137/22M1525004

Karvonen T. (2023). Small sample spaces for Gaussian processes. *Bernoulli: Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, *29*(2), 875–900. https://doi.org/10.3150/22-BEJ1483

Karvonen T., Oates C. J., & Särkkä S. (2018). A Bayes–Sard cubature method. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*.

Karvonen T., Wynne G., Tronarp F., Oates C., & Sarkka S. (2020). Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions. *SIAM/ASA Journal on Uncertainty Quantification*, *8*(3), 926–958. https://doi.org/10.1137/20M1315968

Kennedy M. C., & O'Hagan A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, *87*(1), 1–13. https://doi.org/10.1093/biomet/87.1.1

Larkin F. (1967). Some techniques for rational interpolation. *The Computer Journal*, *10*(2), 178–187. https://doi.org/10.1093/comjnl/10.2.178

Lemaire V., & Pagès G. (2017). Multilevel Richardson–Romberg extrapolation. *Bernoulli: Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, *20*(3), 1029–1067. https://doi.org/10.3150/16-BEJ822

Lucia D. J., Beran P. S., & Silva W. A. (2004). Reduced-order modeling: New approaches for computational physics. *Progress in Aerospace Sciences*, *40*(1-2), 51–117. https://doi.org/10.1016/j.paerosci.2003.12.001

Majda A. J., & Gershgorin B. (2010). Quantifying uncertainty in climate change science through empirical information theory. *Proceedings of the National Academy of Sciences*, *107*(34), 14958–14963. https://doi.org/10.1073/pnas.1007009107

Peherstorfer B., Willcox K., & Gunzburger M. (2018). Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review: Society for Industrial and Applied Mathematics*, *60*(3), 550–591. https://doi.org/10.1137/16M1082469

Piperni P., DeBlois A., & Henderson R. (2013). Development of a multilevel multidisciplinary-optimization capability for an industrial environment. *AIAA Journal: American Institute of Aeronautics and Astronautics*, *51*(10), 2335–2352. https://doi.org/10.2514/1.J052180

Rasmussen C. E., & Williams C. K. (2006). *Gaussian processes for machine learning*. Springer.

Rhee C.-H., & Glynn P. W. (2015). Unbiased estimation with square root convergence for SDE models. *Operations Research*, *63*(5), 1026–1043. https://doi.org/10.1287/opre.2015.1404

Richardson L. F. (1911). The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society A*, *210*(459–470), 307–357. https://doi.org/10.1098/rsta.1911.0009

Richardson L. F., & Gaunt J. A. (1927). The deferred approach to the limit. *Philosophical Transactions of the Royal Society A*, *226*, 223–361. https://doi.org/10.1098/rsta.1927.0008

Sacks J., Welch W. J., Mitchell T. J., & Wynn H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, *4*(4), 409–423. https://doi.org/10.1214/ss/1177012413

Shanks D. (1955). Non-linear transformations of divergent and slowly convergent sequences. *Journal of Mathematics and Physics*, *34*(1-4), 1–42. https://doi.org/10.1002/sapm.v34.1

Sidi A. (2003). *Practical extrapolation methods: Theory and applications*. Cambridge University Press.

Stein M. L., & Hung Y. (2019). Comment on "Probabilistic integration: A role in statistical computation?". *Statistical Science*, *34*(1), 34–37. https://doi.org/10.1214/18-STS677

Strocchi M., Gsell M. A., Augustin C. M., Razeghi O., Roney C. H., Prassl A. J., Vigmond E. J., Behar J. M., Gould J. S., Rinaldi C. A., Bishop M. J., Plank G., & Niederer S. A. (2020). Simulating ventricular systolic motion in a four-chamber heart model with spatially varying robin boundary conditions to model the effect of the pericardium. *Journal of Biomechanics*, *101*, 109645. https://doi.org/10.1016/j.jbiomech.2020.109645

Strocchi M., Longobardi S., Augustin C. M., Gsell M. A., Petras A., Rinaldi C. A., Vigmond E. J., Plank G., Oates C. J., Wilkinson R. D., & Niederer S. A. (2023). Cell to whole organ global sensitivity analysis on a four-chamber heart electromechanics model using Gaussian processes emulators. *PLoS Computational Biology*, *19*(6), e1011257. https://doi.org/10.1371/journal.pcbi.1011257

Teymur O., Foley C., Breen P., Karvonen T., & Oates C. J. (2021). Black box probabilistic numerics. In *Proceedings of the 35th Conference on Neural Information Processing Systems*.

Thiele T. N. (1909). *Interpolationsrechnung*. BG Teubner.

Thodoroff P., Kaiser M., Williams R., Arthern R., Hosking S., Lawrence N., Byrne J., & Kazlauskaite I. (2023). Multi-fidelity experimental design for ice-sheet simulation. arXiv preprint arXiv:2307.08449.

Tuo R., Wu C. J., & Yu D. (2014). Surrogate modeling of computer experiments with different mesh densities. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, *56*(3), 372–380. https://doi.org/10.1080/00401706.2013.842935

Wendland H. (2004). *Scattered data approximation*. Cambridge University Press.