**Noname manuscript No.**
(will be inserted by the editor)

# A Radial Basis Function Method for Noisy Global Optimisation

**Dirk Banholzer · Jörg Fliege · Ralf Werner**

Received: date / Accepted: date

**Abstract** We present a novel response surface method for global optimisation of an expensive and noisy (black-box) objective function, where error bounds on the deviation of the observed noisy function values from their true counterparts are available. The method is based on the well-established RBF method by Gutmann (2001a,c) for minimising an expensive and deterministic objective function, which has become popular both from a theoretical and practical perspective. To construct suitable radial basis function approximants to the objective function and to determine new sample points for successive evaluation of the expensive noisy objective, the method uses a regularised least-squares criterion. In particular, new points are defined by means of a target value, analogous to the original RBF method. We provide essential convergence results, and address some details regarding the numerical implementation of the method.

**Keywords** Global optimisation · expensive noisy objective function · controlled noise · response surface methods · radial basis functions · approximation

## 1 Introduction

In this paper, we are concerned with solving problems of the form

$$\min_{x \in \mathcal{X}} f(x), \tag{1}$$

Dirk Banholzer
Department of Mathematical Sciences, University of Southampton
Southampton, SO17 1BJ, UK
E-mail: db1f12@southamptonalumni.ac.uk

Jörg Fliege
Department of Mathematical Sciences, University of Southampton
Southampton, SO17 1BJ, UK
E-mail: J.Fliege@soton.ac.uk

Ralf Werner
Institut für Mathematik, Universität Augsburg
86159 Augsburg, Germany
E-mail: ralf.werner@math.uni-augsburg.de

where $\mathcal{X} \subset \mathbb{R}^d$ is a nonempty compact set, and $f : \mathcal{X} \to \mathbb{R}$ is a continuous potentially nonconvex objective function that is expensive to evaluate. We assume that evaluations of the objective function $f$ are perturbed by additive noise, where the level of noise can be controlled by means of pointwise error bounds. Specifically, given that some noisy function values $\hat{f}(x_i)$ are observed at the sample points $x_i \in \mathcal{X}$, $i \in \mathbb{N}$, we will consider two models for noise:

1. the errors between $\hat{f}(x_i)$ to the true but unknown counterparts $f(x_i)$ can be quantified by
$$\left|f(x_i) - \hat{f}(x_i)\right| \leq \epsilon_i, \quad i = 1, \dots, n. \tag{2}$$
for some positive values $\epsilon_i = \epsilon(x_i)$. We denote this case as the case of *fixed noise*, i.e. we have been given an error function $\epsilon : \mathcal{X} \longrightarrow \mathbb{R}_+$ which we can evaluate.
2. we consider the case in which we can improve the error bounds during the course of an iteration, i.e. we consider the $x_1, \dots, x_n$ as iterates of an optimisation algorithm and presume error bounds of the form
$$\left|f(x_i) - \hat{f}^{(n)}(x_i)\right| \leq \epsilon_i^{(n)}, \quad i = 1, \dots, n. \tag{3}$$
for some positive errors $\epsilon_i^{(n)} = \epsilon(x_i, n)$. We denote this case as the case of *iterative noise*. Iterative noise typically occurs if a function evaluation that computes $\hat{f}^{(n)}(x_i)$ uses e.g. a Monte Carlo simulation to evaluate some integral occurring in the definition of the function $f$. Given sufficient computational budget, it is then possible to improve a previously computed estimate $\hat{f}^{(k)}(x_i)$ in a later iteration $n > k$ to a hopefully better estimate $\hat{f}^{(n)}(x_i)$ by increasing the sample size of the Monte Carlo simulation. In this paper, we will be concerned with *vanishing iterative noise* in which the $\epsilon_i^{(n)}$ converge to zero for $n, i \to \infty$ in some form.

To clarify terminology and avoid any confusion for the purpose of this paper, we define noise to be any inaccuracy in the function evaluation of $f$. In view of problem (1), this is then sometimes also referred to as minimising a noisy objective function $\hat{f}$ on the parameter space $\mathcal{X}$, see, e.g., Kelley (1999). For notational convenience, from now on we always write $\hat{f}^{(n)}$, where in the case of fixed noise we interpret this as $\hat{f}^{(n)} \equiv \hat{f}$.

To effectively tackle the minimisation of a nonconvex and expensive (black-box) objective function, *response surface methods* have been developed. Their basic idea is to approximate the underlying objective function by a sequence of response surface models, i.e. approximants to the function $f$, that guide the selection of new evaluation points to eventually find a global optimum of the original function. To remain easy to handle and cheap to evaluate, the response surface models are usually composed of simple basis functions and fit to the unknown objective function at a limited number of points, either through interpolation or some approximation scheme. Based on the models, new evaluation points are then iteratively determined by some strategy, which ideally balances between selecting points in unexplored regions of the domain to improve the accuracy of the models there, i.e. a global search, and trusting the models in regions with many function evaluations to find a minimum thereof, i.e. a local search. In this way, the models are successively refined to capture the global behaviour of the objective function as best as possible.

Within the class of response surface methods, various methods can be distinguished, see, e.g., Jones (2001) or, more recently, Vu et al. (2017) for a comparative survey, and Forrester and Keane (2009) for a more practical overview. Very generally, one may say that there are three main methodologies according to which traditional response surface methods may be classified. The underlying idea of *Bayesian methods* is to interpret the objective function $f$ as a realisation of a stochastic process $F : \mathcal{X} \times \Omega \to \mathbb{R}$, $(x, \omega) \mapsto F(x, \omega)$, on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that upon observing $f(x_1), \dots, f(x_n)$, the conditional mean function $s_n(x) = \mathbb{E}_{\mathbb{P}}[F(x)|F(x_1) = f(x_1), \dots, F(x_n) = f(x_n)]$ and the variance function $v_n(x) = \mathbb{V}\mathrm{ar}_{\mathbb{P}}(F(x)|F(x_1) = f(x_1), \dots, F(x_n) = f(x_n))$ act as a response surface and a measure of the involved error to $f$, respectively. In particular, if the underlying stochastic process is assumed to be Gaussian with mean function $\mathbb{E}_{\mathbb{P}}[F(x)]$ and covariance function $\mathbb{C}\mathrm{ov}_{\mathbb{P}}(F(x), F(y))$, $x, y \in \mathcal{X}$, the distributional properties imply that the conditional process is again Gaussian, allowing for suitable expressions in closed form. As for determining new evaluation points, two main strategies can essentially be distinguished in a Bayesian framework. The first strategy, known as P-algorithm, dates back to Kushner (1962, 1964) and maximises the probability of achieving a certain target function value below the current minimum of the surface in order to find a new point. The second strategy has its origin in Mockus et al. (1978) and determines a new point by maximising the expected improvement over the current best function value.

Similar to Bayesian methods, *regression-based methods*, which are commonly referred to as Kriging (Matheron (1963)), also assume $f$ to be a realisation of a stochastic process $\{F(x)\}_{x \in \mathcal{X}}$ but use a linear regression to fit a response surface model. Specifically, given the observations $f(x_1), \dots, f(x_n)$, these methods derive a response surface $s_n$ as the best linear unbiased predictor and the corresponding error $v_n$ as the mean squared error, see, e.g., Sacks et al. (1989) for more details, thus leading to the same methods as in the Bayesian methodology for the special case of a Gaussian process (see, e.g., Fowkes (2011) for the equivalence), but otherwise different ones. The most popular method embedded within a regression-based methodology is the Efficient Global Optimisation (EGO) algorithm by Jones et al. (1998), which specifies the covariance structure of the stochastic process between sampled points by a Gaussian correlation function and finds the next evaluation point by using the expected improvement criterion, as suggested for Bayesian methods. Schonlau (1997) observes that the latter criterion, being independent of any parameter, may result in a search that overly emphasises local search and suggests using a generalised expected improvement, which introduces an additional parameter that controls the balance between global versus local search. Another related modification that allows to exogenously control local and global search by an additional parameter is the weighted expected improvement, due to Sóbester et al. (2005). The EGO approach to construct suitable response surfaces has also been used by Villemonteix et al. (2009) in their Informational Approach to Global Optimisation (IAGO). However, instead of expected improvement, they use the conditional entropy of a minimiser as a criterion to iteratively determine new evaluation points. Further work allow for constraints Habib et al. (2016), multiobjective problems Feliot et al. (2017), and parallelization Wang et al. (2020).

For methods that do not model the objective function by means of stochastic processes, a *general response surface technique* for finding a new evaluation point is proposed by Jones (1996). It assumes the existence of a linear space of functions $\mathcal{A}$,

which is left unspecified but admits a measure of 'bumpiness' $\sigma(s)$ for its elements $s \in \mathcal{A}$. In any given iteration, once a response surface model has been fitted to a set of function values $f(x_1), \ldots, f(x_n)$ through interpolation, a target value $f^*$ is then chosen which may be considered as a rough estimate of the global minimum of $f$. By this choice, a new evaluation point $x_{n+1}$ is then determined as the value of $y \notin \{x_1, \ldots, x_n\}$ such that the augmented response surface $s_y \in \mathcal{A}$ minimises the bumpiness $\sigma(s)$ on $\mathcal{A}$, subject to the interpolation conditions

$$\begin{aligned} s_y(x_i) &= f(x_i), \qquad i = 1, \ldots, n, \\ s_y(y) &= f^*, \end{aligned}$$

and provided that, for any $y \in \mathcal{X} \backslash \{x_1, \ldots, x_n\}$, the interpolant $s_y \in \mathcal{A}$ is uniquely defined. The most prominent response surface method that is based on Jones' general technique is suggested by Gutmann (2001a,c) in form of the radial basis function (RBF) method. As the name suggests, the method relies on the use of radial basis functions, which not only ensures the uniqueness of interpolants under relatively mild conditions on the location of the sample points, but also provides in a natural way a measure of 'bumpiness' in form of a semi-norm. The strategy for determining new evaluation points is based on a mathematically sound mechanism, which facilitates establishing convergence of the method and its close relation to the P-algorithm. On the practical side, the method has proven to be a powerful tool and performs well on well-behaved expensive optimisation problems, see, e.g., Björkman and Holmström (2000), while the noted slow convergence of the method to a global minimum of more complex objective functions has been addressed by Regis and Shoemaker (2007b), Holmström (2008), and Cassioli and Schoen (2013), most notably. In addition, Costa and Nannicini (2018) propose a technique to speed up the practical convergence of Gutmann's RBF method in case noisy and less expensive function values are additionally available. Other works include the investigation of multistart methods Regis and Shoemaker (2013a), the multiobjective case Akhtar and Shoemaker (2016); Regis (2016); Wang et al. (2022), and the case of black-box constraints Regis (2014). The case of noisy objective functions has only recently been considered in Shen and Shoemaker (2020).

Finally, we would like to mention further approaches which seemingly do not rely on any of the described underlying methodologies and are thus designed to work with any kind of response surface model, see, e.g, Regis and Shoemaker (2005, 2007a,c, 2013b) and Ji et al. (2013).

Despite its importance in applications, the global optimisation of expensive objective functions in the presences of noise has attracted considerably less attention than the equivalent optimisation without noise. Due to the underlying probabilistic framework that is provided by Bayesian and regression-based methods, the fundamental derivations of response surfaces and corresponding error measures, i.e. of the conditional mean and variance functions as well as of the unbiased predictors and mean squared errors, respectively, can be extended straightforwardly to noisy observations, by adding randomly distributed error terms to the modelling stochastic process, see, e.g., Schonlau (1997). Yet, the determination of new evaluation points poses a substantial difficulty, which has been dealt with differently by a few authors. In their Sequential Kriging Optimisation (SKO) method, Huang et al. (2006) extend the EGO algorithm to noisy objective function values, assuming that the involved random errors are i.i.d. normally distributed with

constant variance. Correspondingly, the method relies on the same Gaussian covariance structure of the EGO algorithm plus a variance term, and to select new evaluation points, an augmented expected improvement criterion is derived that calculates a scaled expected improvement over the response surface value of the so-called effective best solution instead of the current minimum function value. A further extension to the expected improvement criterion that may be used in a Bayesian/regression-based setup for noisy objective function values is suggested by Gramacy and Lee (2011), also known as the integrated expected conditional improvement. Their main idea is to consider the expected improvement at a reference point, given that the objective function has been sampled at a candidate point, i.e. the expected conditional improvement, and find the next point as the maximiser of this criterion, integrated by a suitable density function over all reference points. Finally, in Villemonteix et al. (2009), the authors also show that their IAOG method for exact function values can be extended to handle noisy observations. Specifically, they assume that the errors in the observed function values are i.i.d. normally distributed with known mean and variance, and estimate the conditional entropy criterion by simulating on the noisy observed function values, instead of the true ones.

Relating to Jones' general technique in the presence of noise, Žilinskas (2010) addresses the similarity between the P-algorithm and Gutmann's RBF method and shows that these techniques can be extended to noisy function values if appropriate modifications are made in both algorithms. In particular, he suggests to construct radial basis function approximants for the latter by minimising the semi-norm such that the residual sum of squares of an approximant to the noisy observations is proportional to the variance of the involved additive noise, which is assumed to be constant and known. New evaluation points may then be determined similar to Jones' technique by means of target values, i.e. by minimising the semi-norm of an augmented surface such that it interpolates a chosen target value and such that the residual sum of squares of the augmented surface to the noisy observations is proportional to the known variance. However, even though Žilinskas establishes the theoretical similarity between the P-algorithm and the RBF method in a noisy setup, no explicit algorithm making use of this result is proposed. Radial basis functions are also used in the algorithm by Jakobsson et al. (2010), called qualSolve, for the global optimisation of expensive black-box functions subject to noise. Here, the authors construct response surfaces by minimising the convex sum of the squared semi-norm of a radial basis function approximant and the squared difference between its values at the sample points and the noisy observations, where an additionally introduced parameter to balance between both measures is estimated by cross validation. Moreover, to select new evaluation points a quality function is maximised, which is calculated at each point by the minimum distance to previously evaluated points and weighted by the response surface value at that point. In particular, the weights are adjusted periodically in order to alternate between local and global search of the method.

Given above contributions, we present in this paper a novel RBF method for noisy objective functions in which the level of noise can be controlled by means of pointwise error bounds. The method is essentially based on Gutmann's original RBF method for deterministic objective functions and uses some of the ideas from Žilinskas (2010) to extending Gutmann's method to a noisy setup. In establishing the method, we address the construction of appropriate response surfaces and

the determination of new evaluation points once a surface has been constructed, as these are the two main components that require specification in order to deal with noisy function values. In particular, since radial basis function interpolation is no longer feasible in the present situation, we first consider common approaches for the approximation of a noisy function by means of radial basis functions and briefly discuss their suitability for integration into a response surface method. As regularised least-squares approximants explicitly seek to balance between the bumpiness of the surface and the closeness to the data, where the additional regularisation parameter may be set in accordance with the available error bounds, they turn out to be particularly suited for our purposes. Moreover, the least-squares criterion allows for a convenient adaption of Jones' technique to determine new evaluation points through target values, by analogy with Gutmann's original algorithm. In particular, this functionality then also allows to establish convergence of the method, where we show that the convergence properties of Gutmann's deterministic method are kept when the exact function values are replaced by corresponding noisy values. As we will see, convergence can be achieved by updating the regularisation parameter in a particular way, depending on the model of noise:

1. in the case of fixed noise, it is sufficient to ensure that the sequence of regularisation parameters converges to zero quickly enough. As such, it suffices to choose these parameters according to some exogenous sequence, at least for theoretical purposes.
2. in the case of iterative noise, if the noise vanishes fast enough, it is possible to choose the regularisation parameter in each step in such a way that the bumpiness of the approximant is as small as possible, thereby greatly simplifying the inner optimisation step in which an augmented function based on this approximant is minimised.

The remainder of this paper is organised as follows. In Section 2, we review Gutmann's original RBF method to minimise a deterministic nonconvex objective function that is expensive to evaluate. In Section 3, we briefly outline common approaches for radial basis function approximation and discuss their suitability for integration into a response surface method. Based on regularised least-squares approximants, we then present in Section 4 a RBF method for minimising a noisy nonconvex and expensive objective function, given that error bounds on the observed function values are available. In Section 5, we establish the convergence of the method. Finally, Section 6 contains our conclusions.

## 2 Gutmann's RBF method

Let us briefly describe each step of Gutmann's original RBF method (2001a; 2001c) for deterministic objective functions, as this will provide us with the necessary tools to generalise this method to the noisy case considered in this paper. The method relies on the general technique by Jones (1996), but specifically employs radial functions to construct response surface interpolants of the generic form

$$s(x) = \sum_{i=1}^{n} \lambda_i \phi(\|x - x_i\|_2) + p(x), \quad x \in \mathbb{R}^d, \tag{4}$$

where $\phi : [0,\infty) \to \mathbb{R}$ is a fixed radial function, $\{\lambda_i\}_{i=1}^n$ are real coefficients, $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ are distinct centre points, and $p \in \mathcal{P}_m^d$ is a polynomial from the linear space of all real-valued polynomials of total degree at most $m-1$ in $d$ variables, with $\mathcal{P}_0^d = \{0\}$. On the linear space of all functions of the form (4) on $\mathcal{X}$, formally defined by $\mathcal{A}_{\phi,m}(\mathcal{X}) := \mathcal{F}_\phi(\mathcal{X}) + \mathcal{P}_m^d$ with

$$\mathcal{F}_\phi(\mathcal{X}) := \Big\{ \sum_{i=1}^n \lambda_i \phi(\|\cdot - x_i\|_2) : n \in \mathbb{N}, \lambda \in \mathbb{R}^n, \{x_i\}_{i=1}^n \subset \mathcal{X}, \sum_{i=1}^n \lambda_i p(x_i) = 0,\ p \in \mathcal{P}_m^d \Big\}, \tag{5}$$

a measure of 'bumpiness' is then given in a natural way by the semi-norm $\|\cdot\|_\phi := \langle \cdot,\cdot \rangle_\phi^{1/2}$, induced by the semi-inner product

$$\langle s, u \rangle_\phi := \sum_{i=1}^{n(s)} \lambda_i^s u(x_i^s), \tag{6}$$

for any two elements $s, u \in \mathcal{A}_{\phi,m}(\mathcal{X})$ with

$$s(x) = \sum_{i=1}^{n(s)} \lambda_i^s \phi(\|x - x_i^s\|_2) + p^s(x) \quad \text{and} \quad u(x) = \sum_{i=1}^{n(u)} \lambda_i^u \phi(\|x - x_i^u\|_2) + p^u(x).$$

The classical choices of radial basis functions $\phi$, along with their minimal order $m_\phi$ guaranteeing conditional positive definiteness are given in Table 1, cf. Gutmann (2001a), listing (2.2), or Gutmann (2001c), listing (3.2). In what follows, we will also need the notion of *conditionally positive definite functions*. Recall that a continuous radial function $\phi$ is conditionally positive definite of order $m$ if $\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \phi(\|x_i - x_j\|_2) > 0$ for any pairwise distinct points $x_1, \ldots, x_n$, $n \in \mathbb{N}$, and any $\lambda \in \mathbb{R}^n \backslash \{0\}$ satisfying $\sum_{i=1}^n \lambda_i p(x_i) = 0$, $p \in \mathcal{P}_m^d$. Since a conditionally positive definite function of order $m_1$ is also conditionally positive definite of order $m_2 \geq m_1$, particular interest is given to the smallest possible order $m_\phi \in \mathbb{N}_0$ such that $\phi$ is conditionally positive definite.

| Radial basis function | $\phi(r)$ | Specification | Minimal order $m_\phi$ |
|---|---|---|---|
| Surface splines | $r^\nu$<br>$r^\nu \log r$ | $\nu \in \mathbb{N}$, $\nu$ odd<br>$\nu \in \mathbb{N}$, $\nu$ even | $\lfloor \nu/2 \rfloor + 1$ |
| Multiquadrics | $(r^2 + \zeta^2)^\nu$ | $\nu > 0$, $\nu \notin \mathbb{N}$ | $\lfloor \nu \rfloor + 1$ |
| Inverse multiquadrics | $(r^2 + \zeta^2)^\nu$ | $\nu < 0$ | 0 |
| Gaussians | $e^{-\zeta r^2}$ | | 0 |

Table 1: Common choices of radial basis functions, their shape parameter $\zeta > 0$, smoothing parameter $\nu$, and the minimal order $m_\phi$.

*Construction of a Response Surface*

Suppose we are in iteration $n$ of our algorithm and can now interpolate the data $(x_1, f(x_1)), \dots, (x_n, f(x_n))$. To construct an interpolant $s_n \in \mathcal{A}_{\phi,m}(\mathcal{X})$ of the form (4), with $\mathcal{A}_{\phi,m}(\mathcal{X})$ defined as in (5), the corresponding coefficients are determined by solving

$$\min_{s \in \mathcal{A}_{\phi,m}(\mathcal{X})} \|s\|_\phi \quad \text{s.t.} \quad s(x_i) = f(x_i), \quad i = 1, \dots, n, \tag{7}$$

which reduces to solving the linear system (see Schaback Schaback (1993))

$$\begin{pmatrix} \Phi & P \\ P^\top & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ c \end{pmatrix} = \begin{pmatrix} F \\ 0 \end{pmatrix}, \tag{8}$$

where $\Phi \in \mathbb{R}^{n \times n}$ and $P \in \mathbb{R}^{n \times \widetilde{m}}$ denote the interpolation and polynomial basis matrix with entries $\Phi_{ij} = \phi(\|x_i - x_j\|_2)$, $i, j = 1, \dots, n$ and $P_{ij} = p_j(x_i)$, $i = 1, \dots, n$, $j = 1, \dots, \widetilde{m}$, respectively, $\lambda \in \mathbb{R}^n$ and $c \in \mathbb{R}^{\widetilde{m}}$ are the coefficient vectors, and $F = (f(x_1), \dots, f(x_n))^\top$ stands for the vector of observed function values.

The unique solvability of the linear system (8) follows under the relatively mild condition that the sample points $x_1, \dots, x_n$ form a $\mathcal{P}_m^d$-unisolvent set, i.e. if $p \in \mathcal{P}_m^d$ and $p(x_i) = 0$, $i = 1, \dots, n$, then $p \equiv 0$, see, e.g., Wendland (2005b) for details. Moreover, it is easy to verify that the linear system will remain uniquely solvable upon the successive addition of new data points, provided that they are distinct from previous ones.

*Determination of the Next Evaluation Point*

Upon the construction of $s_n$, the next evaluation point $x_{n+1}$ is determined according to Jones' general technique. For a given target value $f_n^*$ that will be specified further below, the point $x_{n+1}$ is given as the point $y \in \mathcal{X} \backslash \{x_1, \dots, x_n\}$ such that there is an *augmented surface* $s_y \in \mathcal{A}_{\phi,m}(\mathcal{X})$ that solves

$$\begin{aligned} \min_{\substack{y \in \mathcal{X} \\ s \in \mathcal{A}_{\phi,m}(\mathcal{X})}} \quad & \|s\|_\phi \quad \text{s.t.} \quad s(x_i) = f(x_i), \quad i = 1, \dots, n, \\ & s(y) = f_n^*. \end{aligned} \tag{9}$$

To simplify problem (9), the optimal interpolant $s_y \in \mathcal{A}_{\phi,m}(\mathcal{X})$, $y \in \mathcal{X} \backslash \{x_1, \dots, x_n\}$, satisfying the interpolation conditions in (9) can be rewritten as

$$s_y(x) = s_n(x) + \big[f_n^* - s_n(y)\big] l_n(y, x), \quad x \in \mathbb{R}^d, \tag{10}$$

where $l_n(y, \cdot) \in \mathcal{A}_{\phi,m}(\mathcal{X})$ is the optimal interpolant to

$$\begin{aligned} l_n(y, x_i) &= 0, \qquad i = 1, \dots, n, \\ l_n(y, y) &= 1. \end{aligned} \tag{11}$$

In particular, the function $l_n(y, \cdot)$ can be expressed as

$$l_n(y, x) = \sum_{i=1}^{n} \alpha_i(y) \phi(\|x - x_i\|_2) + \beta(y) \phi(\|x - y\|_2) + \sum_{j=1}^{\widetilde{m}} b_j(y) p_j(x), \quad x \in \mathbb{R}^d,$$

whose coefficients $\alpha(y) = (\alpha_1(y), \dots, \alpha_n(y))^\top \in \mathbb{R}^n$, $\beta(y) \in \mathbb{R}$ and $b(y) = (b_1(y), \dots, b_{\widetilde{m}}(y))^\top \in \mathbb{R}^{\widetilde{m}}$ solve the linear system

$$\begin{pmatrix} \Phi & u_n(y) & P \\ u_n(y)^\top & \phi(0) & \pi(y)^\top \\ P^\top & \pi(y) & 0 \end{pmatrix} \begin{pmatrix} \alpha(y) \\ \beta(y) \\ b(y) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \tag{12}$$

for the matrices $\Phi \in \mathbb{R}^{n \times n}$ and $P \in \mathbb{R}^{n \times \widetilde{m}}$, and the vectors $u_n(y) := (\phi(\|x_1 - y\|_2), \dots, \phi(\|x_n - y\|_2))^\top \in \mathbb{R}^n$ and $\pi(y) := (p_1(y), \dots, p_{\widetilde{m}}(y))^\top \in \mathbb{R}^{\widetilde{m}}$. By means of representation (10), the squared semi-norm of $s_y$ can then be simplified to

$$\begin{aligned} \|s_y\|_\phi^2 &= \|s_n\|_\phi^2 + 2\big[f_n^* - s_n(y)\big]\langle s_n, l_n(y, \cdot)\rangle_\phi + \big[f_n^* - s_n(y)\big]^2 \|l_n(y, \cdot)\|_\phi^2 \\ &= \|s_n\|_\phi^2 + \beta(y)\big[f_n^* - s_n(y)\big]^2, \end{aligned} \tag{13}$$

using the definition of the semi-inner product (6) and the interpolation conditions (11). Since $\|s_n\|_\phi$ is independent of $y$, equation (13) shows that the required minimisation of $\|s_y\|_\phi$ with respect to $y$ boils down to minimising the nonnegative function

$$g_n(y) := \mu_n(y)\big[f_n^* - s_n(y)\big]^2, \qquad y \in \mathcal{X}\backslash\{x_1, \dots, x_n\}, \tag{14}$$

where the function $\mu_n : \mathcal{X}\backslash\{x_1, \dots, x_n\} \to \mathbb{R}$ is given by

$$\mu_n(y) := \|l_n(y, \cdot)\|_\phi^2 = \beta(y). \tag{15}$$

Note that the function $\mu_n$ is well-defined and allows for the properties described in the following two remarks.

*Remark 1* Definition (15) provides that $\mu_n(y) > 0$ for $y \in \mathcal{X}\backslash\{x_1, \dots, x_n\}$: Assuming there is an $y_0 \in \mathcal{X}\backslash\{x_1, \dots, x_n\}$ with $\mu_n(y_0) = 0$, definition (15) and the $\mathcal{P}_m^d$-unisolvency of $\{x_1, \dots, x_n\}$ yield $l_n(y_0, \cdot) \equiv 0$. This, however, is in contradiction to the interpolation constraint $l_n(y_0, y_0) = 1$.

*Remark 2* By applying Cramer's rule to the linear system (12), the function $\mu_n$ can be computed as

$$\mu_n(y) = \frac{\det A_n}{\det A_n(y)}, \qquad y \in \mathcal{X}\backslash\{x_1, \dots, x_n\},$$

where $A_n$ and $A_n(y)$ are given by the nonsingular interpolation matrices on the left-hand sides of equations (8) and (12), respectively. Hence, since $\det A_n$ is a nonzero constant and $\lim_{y \to x_i} \det A_n(y) = 0$ for any $i \in \{1, \dots, n\}$, it follows that

$$\lim_{y \to x_i} \mu_n(y) = \infty, \qquad i = 1, \dots, n.$$

*Choice of Target Value*

The choice of $f_n^*$ crucially influences the location of the new point $x_{n+1}$. To guarantee that $x_{n+1}$ as a global minimiser of $g_n$ on $\mathcal{X}\backslash\{x_1,\dots,x_n\}$ exists and does not coincide with previous sample points, it must hold that

$$f_n^* \in \big[-\infty, \min_{y\in\mathcal{X}} s_n(y)\big], \tag{16}$$

where the case $f_n^* = \min_{y\in\mathcal{X}} s_n(y)$ is only admissible if $f_n^* < s_n(x_i)$, $i = 1,\dots,n$.

Specifically, for low target values satisfying (16), the method essentially performs a global search in which the new point $x_{n+1}$ is sampled away from already evaluated points. A high target value close or equal to $\min_{y\in\mathcal{X}} s_n(y)$ is supposed to sample $x_{n+1}$ either in the vicinity of a global minimiser of $s_n$, if $f_n^* < \min_{y\in\mathcal{X}} s_n(y)$, or as a global minimiser of $s_n$, if $f_n^* = \min_{y\in\mathcal{X}} s_n(y)$, cf. Regis and Shoemaker (2007b). In particular, for $f_n \to -\infty$, one can observe by the definition of $g_n$ and the boundedness of $s_n$ on $\mathcal{X}$ that $\mu_n(x_{n+1}) \le \mu_n(y)$, $y \in \mathcal{X}\backslash\{x_1,\dots,x_n\}$. Hence, choosing $f_n^* = -\infty$ reduces the minimisation of $g_n$ on $\mathcal{X}\backslash\{x_1,\dots,x_n\}$ to the minimisation of $\mu_n$, which samples $x_{n+1}$ as far away as possible from the points $x_1,\dots,x_n$.

*Summary of Gutmann's RBF Method*

Altogether, Gutmann's RBF method for minimising a deterministic and continuous function $f : \mathcal{X} \to \mathbb{R}$ on a compact set $\mathcal{X}$ can be summarised as follows.

**Algorithm 1** *(Gutmann's RBF Method).*

0. ***Initial step:***
   - *Choose a conditionally positive definite radial basis function $\phi$ of order $m$.*
   - *Generate a $\mathcal{P}_m^d$-unisolvent set of points $\{x_1,\dots,x_{n_0}\} \subset \mathcal{X}$.*
   - *Evaluate $f$ at the points $x_1,\dots,x_{n_0}$, and set $n = n_0$.*

1. ***Iteration step:***
   ***while*** $n \le n^{\max}$ ***do***
   - *Construct the interpolant $s_n \in \mathcal{A}_{\phi,m}(\mathcal{X})$ solving*

$$\min_{s\in\mathcal{A}_{\phi,m}(\mathcal{X})} \|s\|_\phi \qquad s.t. \quad s(x_i) = f(x_i), \quad i = 1,\dots,n.$$

   - *Choose an admissible target value $f_n^* \in \big[-\infty, \min_{y\in\mathcal{X}} s_n(y)\big]$.*
   - *Determine $x_{n+1}$, which is the value of $y$ that solves*

$$\min_{y\in\mathcal{X}\backslash\{x_1,\dots,x_n\}} \mu_n(y)\big[f_n^* - s_n(y)\big]^2.$$

   - *Evaluate $f$ at $x_{n+1}$, and set $n = n+1$.*

   ***end while***

## 3 Approximation with Radial Basis Functions

To recover an unknown function $f : \mathcal{X} \to \mathbb{R}$ on some set $\mathcal{X} \subset \mathbb{R}^d$ from a number of observed function values $f(x_1), \ldots, f(x_n)$ with $x_1, \ldots, x_n \in \mathcal{X}$, an interpolation technique is typically adopted if the respective function values are known to be exact. However, if the observations are contaminated by noise, i.e. we observe $\hat{f}^{(n)}(x_1), \ldots, \hat{f}^{(n)}(x_n)$ in step $n$ of our algorithm, then other approximation techniques are recommended. In particular, in such cases, too much weight would be given to the involved noise, which may easily lead to a model overfitting the data and becoming unnecessarily oscillating, thus corresponding poorly to the underlying function.

Unlike in the case of interpolation, there exist various possibilities to approximate a set of noisy function values by means of radial basis functions, where a suitable choice essentially depends on the nature of the available data and the intended use of the resulting approximant. A technique that is frequently employed is a *least-squares approximation*, see, e.g., Buhmann (2003), Chapter 8, or Iske (2004), Section 3.10, where approximants of the generic form (4) are considered for a reduced numbered of pairwise distinct centres $\{\tilde{x}_j\}_{j=1}^{\tilde{n}} \subset \mathcal{X}$, $\tilde{n} + \widetilde{m} < n$, which usually coincide with some of the sample points $x_1, \ldots, x_n$, but may also be different. This form is then used to obtain an optimal approximant $s \in \tilde{\mathcal{A}}_\phi(\mathcal{X})$ by solving

$$\min_{s \in \tilde{\mathcal{A}}_\phi(\mathcal{X})} \sum_{i=1}^{n} w_i \big(s(x_i) - \hat{f}^{(n)}(x_i)\big)^2, \tag{17}$$

where $\tilde{\mathcal{A}}_\phi(\mathcal{X})$ denotes the corresponding linear function space and $w_1, \ldots, w_n$ are positive weights to take care of potential heteroscedasticity in the data. Due to the side conditions in $\tilde{\mathcal{A}}_\phi(\mathcal{X})$, problem (17) constitutes a linear least-squares problem with equality constraints, which can be solved uniquely via a linear system if the set of centres $\{\tilde{x}_1, \ldots, \tilde{x}_{\tilde{n}}\}$ is $\mathcal{P}_m^d$-unisolvent and forms a subset of the sample points, see, e.g., Iske (2004), Theorem 17. A least-squares approach may notably reduce the complexity of constructing an approximant if $\tilde{n} \ll n$. However, the main drawback then lies in choosing a suitable set of centres which defines both the smoothness of an approximant and its closeness to the data. This ambiguity makes it difficult to incorporate the technique into a response surface method where new points are added iteratively, as argued, for instance, by Žilinskas (2010).

An approach that explicitly allows to include both the semi-norm as a measure of smoothness and the availability of error bounds into the construction of a radial basis function approximant is known as *relaxed interpolation*, see, e.g., Schaback and Wendland (2006), Section 3. Specifically, requiring an approximant to be as smooth as possible but such that it deviates at the sampled points $x_i$ from the observed values $\hat{f}^{(n)}(x_i)$ by at most $\epsilon_i$, an optimal approximant $s \in \mathcal{A}_{\phi,m}(\mathcal{D})$ is found by solving

$$\begin{aligned} \min_{s \in \mathcal{A}_{\phi,m}(\mathcal{X})} \quad & \|s\|_\phi^2 \\ \text{s.t.} \quad & \big|w_i\big(s(x_i) - \hat{f}^{(n)}(x_i)\big)\big| \leq \epsilon_i, \quad i = 1, \ldots, n. \end{aligned} \tag{18}$$

By definition of the semi-norm and the side conditions in $\mathcal{A}_{\phi,m}(\mathcal{X})$, problem (18) presents a convex quadratic programme with both equality and inequality constraints, which can be solved uniquely if the set of points $\{x_1, \ldots, x_n\}$ is assumed

to be $\mathcal{P}_m^d$-unisolvent. Note that as a consequence of the involved inequality constraints, the convex quadratic programme needs to be solved; the optimal approximant can no longer be determined by solving just a linear system of equations. By applying the KKT conditions, it can be shown that the optimal approximant either interpolates the endpoints of the (potentially scaled) error bounds or the corresponding coefficient $\lambda_i$ equals zero, or both, cf. Schölkopf and Smola (2002) for the related concept of support-vector machines.

The *regularised least-squares approximation*, as described, for instance, in Wendland and Rieger (2005) or Wendland (2005a), is another approach that explicitly incorporates the semi-norm into the construction of the approximant. However, instead of imposing inequality constraints to regulate the discrepancy to the noisy function values, the closeness to the data is assessed by residual sum of squares. Consequently, an optimal approximant $s^\gamma \in \mathcal{A}_\phi(\mathcal{X})$ is sought as the solution of

$$\min_{s \in \mathcal{A}_{\phi,m}(\mathcal{X})} \gamma \|s\|_\phi^2 + \frac{1}{n} \sum_{i=1}^{n} w_i \big(s(x_i) - \hat{f}^{(n)}(x_i)\big)^2, \tag{19}$$

where the additional parameter $\gamma > 0$ is introduced to control the trade-off between the smoothness of the approximant and its closeness to the noisy function values. In particular, for large $\gamma$ we place more emphasis on minimizing the bumpiness, while for small $\gamma$ the closeness to the data is enforced, yielding an interpolation of $\hat{f}^{(n)}(x_1), \dots, \hat{f}^{(n)}(x_n)$ in case $\gamma = 0$.

Bearing in mind the constraints on the coefficients $\lambda$ in $\mathcal{A}_{\phi,m}(\mathcal{X})$, problem (19) comprises an equality constrained convex quadratic programme. Hence, similar to plain interpolation, the construction of an approximant can be reduced to solving a (regularised) linear system, see Theorem 1 below. In particular, this implies that errors in function values are taken into account by interpolating some perturbed noisy function values, where the magnitude of the perturbation is governed by the regularisation parameter $\gamma$. Moreover, the parameter $\gamma$ has a clear and intuitive interpretation, which facilitates its determination by means of the available error bounds (2) and also allows for a convenient application of Jones' technique to determine new evaluation points, cf. Section 4. Consequently, regularised least-squares approximation seems to provide the most suitable approach for an extension of Gutmann's RBF method to noise. In what follows, we require the following result.

**Theorem 1** *Let $\phi$ be a conditionally positive definite radial basis function of order $m$, and assume that a $\mathcal{P}_m^d$-unisolvent set of points $\{x_1, \dots, x_n\} \subset \mathcal{D}$ with corresponding noisy function values $\hat{f}^{(n)}(x_1), \dots, \hat{f}^{(n)}(x_n)$ is given. Then, for any $\gamma > 0$, the approximant $s_n^\gamma \in \mathcal{A}_{\phi,m}(\mathcal{D})$ whose coefficients are determined by the linear system*

$$\begin{pmatrix} \Phi + n\gamma W^{-1} & P \\ P^\top & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ c \end{pmatrix} = \begin{pmatrix} \widehat{F} \\ 0 \end{pmatrix}, \tag{20}$$

*where $W = \operatorname{diag}(w_1, \dots, w_n)$ and $\widehat{F} = (\hat{f}^{(n)}(x_1), \dots, \hat{f}^{(n)}(x_n))^\top$, is the unique element of $\mathcal{A}_{\phi,m}(\mathcal{D})$ that solves the regularised least-squares approximation problem* (19).

*Proof* Let $\gamma > 0$ be fixed, and observe that problem (19) can be rewritten as

$$\min_{(\lambda,c)^\top \in \mathbb{R}^{n+\widetilde{m}}} n\gamma \lambda^\top \Phi \lambda + \big\| W^{1/2} (\Phi\lambda + Pc - \widehat{F}) \big\|_2^2 \qquad \text{s.t.} \quad P^\top \lambda = 0. \tag{21}$$

By the conditional positive definiteness of $\phi$, problem (21) is strictly convex. Hence, a unique solution exists if the set of sample points $\{x_1, \ldots, x_n\}$ is $\mathcal{P}_m^d$-unisolvent, guaranteeing that the matrix $P^\top$ has full row rank. Applying the KKT conditions to (21) provides the linear equations

$$(n\gamma\Phi + \Phi^\top W\Phi)\lambda + \Phi^\top WPc + P\upsilon = \Phi^\top W\widehat{F} \tag{22}$$

$$P^\top W(\Phi\lambda + Pc) = P^\top W\widehat{F} \tag{23}$$

$$P^\top\lambda = 0, \tag{24}$$

where $\upsilon \in \mathbb{R}^{\widetilde{m}}$ denotes the Lagrange multiplier for the constraint $P^\top\lambda = 0$. Since $\phi$ is conditionally positive definite, the matrix $\Phi$ is invertible for any $\lambda \in \mathbb{R}^n\backslash\{0\}$ satisfying (24). Thus, multiplying equation (22) by $(\Phi W)^{-1}$ simplifies to

$$(\Phi + n\gamma W^{-1})\lambda + Pc + (\Phi^\top W)^{-1}P\upsilon = \widehat{F},$$

which, by substituting into equation (23), yields $P^\top\Phi^{-1}P\upsilon = 0$. However, since $\{x_1, \ldots, x_n\}$ is $\mathcal{P}_m^d$-unisolvent, the latter implies $\upsilon = 0$, such that we obtain the stated linear system (20). □

Note that the linear system (20) remains uniquely solvable if new points are added, distinct from already sampled points.

## 4 A Radial Basis Function Method for Noisy Objective Functions

In this section, we describe our novel RBF method for minimising a noisy objective function $\hat{f} : \mathcal{X} \to \mathbb{R}$ on a compact set $\mathcal{X}$, which proceeds similar to Algorithm 1 but uses a regularised least-squares approach to construct approximating response surfaces and determine new evaluation points.

Let $\phi$ be a conditionally positive definite radial basis function of order $m$ and $\mathcal{P}_m^d$ be the space of polynomials of degree at most $m-1$ with basis $\{p_j\}_{j=1}^{\widetilde{m}}$. Assume that the initially sampled points $x_1, \ldots, x_{n_0}$ form a $\mathcal{P}_m^d$-unisolvent set and that error bounds $\epsilon_1, \ldots, \epsilon_{n_0}$ and positive weights $w_1, \ldots, w_{n_0}$ are available for the corresponding noisy function values $\hat{f}^{(n_0)}(x_1), \ldots, \hat{f}^{(n_)}(x_{n_0})$. For $n \geq n_0$, a general iteration, consisting of the construction of an approximant and the determination of a new evaluation point by a suitably chosen target value, can then be described as follows, cf. Subsection 4.4 for a complete description of the full algorithm.

### 4.1 Construction of Response Surface

For given data $(x_1, \hat{f}^{(n)}(x_1)), \ldots, (x_n, \hat{f}^{(n)}(x_n))$, weights $w_1, \ldots, w_n$ and $\gamma > 0$ we denote the unique solution of the linear system (20), cf. Theorem 1, by $(\lambda^{(n,\gamma)}, c^{(n,\gamma)})$. Recall that we denote the optimal regularised least-squares approximant to the given data from the space $\mathcal{A}_{\phi,m}(\mathcal{X})$ by

$$s_n^\gamma(x) = \sum_{i=1}^{n} \lambda_i^{(n,\gamma)}\phi(\|x - x_i\|_2) + p^{(n,\gamma)}(x), \quad x \in \mathbb{R}^d. \tag{25}$$

To determine $\gamma_n$ in the $n$-th iteration, we make use of the fact that we are predominantly interested in finding a rather smooth approximant that deviates at most by the error bounds from the noisy function values to recover the underlying function $f$. Accordingly, we first observe that the smoothness of the approximant $s_n^\gamma$ can alternatively be characterised in terms of the parameter $\gamma$, which seems intuitively clear from the formulation (19). More formally, it can be justified by the following Proposition 1, for which we define by $\mathcal{R}(P)$ the range of the polynomial basis matrix $P \in \mathbb{R}^{n \times \widetilde{m}}$.

**Proposition 1** *Let $\phi$ be a conditionally positive definite radial basis function of order $m$, and let $\{x_1, \dots, x_n\} \subset \mathcal{X}$ be a $\mathcal{P}_m^d$-unisolvent set with corresponding noisy function values $\hat{f}^{(n)}(x_1), \dots, \hat{f}^{(n)}(x_n)$. For $\gamma > 0$, let $s^\gamma \in \mathcal{A}_{\phi,m}(\mathcal{X})$ denote the unique solution to the regularised least-squares problem* (19)*. Then, the following holds:*

*a) $\lambda^{(n,\gamma)}$ and $c^{(n,\gamma)}$ depend continuously on $\gamma$.*

*b) The optimal value of* (19) *is concave and monotonically increasing in $\gamma$. In case $(\hat{f}^{(n)}(x_1), \dots, \hat{f}^{(n)}(x_n))^\top \notin \mathcal{R}(P)$, then the optimal value function is strictly monotonically increasing.*

*c) For any fixed noisy function values, the term $\|s^\gamma\|_\phi$ is monotonically decreasing in $\gamma$ and $\frac{1}{n}\sum_{i=1}^n w_i (s^\gamma(x_i) - \hat{f}^{(n)}(x_i))^2$ is monotonically increasing in $\gamma$. If $(\hat{f}^{(n)}(x_1), \dots, \hat{f}^{(n)}(x_n))^\top \notin \mathcal{R}(P)$, then these terms are strictly monotonically decreasing and increasing in $\gamma$, respectively.*

*Proof* Since the associated matrix on the left-hand side of the linear system (20) is nonsingular and depends continuously on $\gamma$, so does its inverse, which establishes statement a).

The first part of statement b) follows directly from the affine structure of the objective function in $\gamma$. For the second part note that it holds $\|s^\gamma\|_\phi^2 \neq 0$ due to the assumption $(\hat{f}^{(n)}(x_1), \dots, \hat{f}^{(n)}(x_n))^\top \notin \mathcal{R}(P)$, see Gutmann (2001b), p. 318.

To show c), let $0 < \gamma < \widetilde{\gamma}$ be fixed. By the optimality of the corresponding minimisers $s^\gamma, s^{\widetilde{\gamma}} \in \mathcal{A}_{\phi,m}(\mathcal{X})$, we then have

$$\gamma\|s^\gamma\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i\big(s^\gamma(x_i) - \hat{f}^{(n)}(x_i)\big)^2 \leq \gamma\|s^{\widetilde{\gamma}}\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i\big(s^{\widetilde{\gamma}}(x_i) - \hat{f}^{(n)}(x_i)\big)^2,$$

and

$$\widetilde{\gamma}\|s^{\widetilde{\gamma}}\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i\big(s^{\widetilde{\gamma}}(x_i) - \hat{f}^{(n)}(x_i)\big)^2 \leq \widetilde{\gamma}\|s^\gamma\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i\big(s^\gamma(x_i) - \hat{f}^{(n)}(x_i)\big)^2.$$

Adding both inequalities, cancelling equal terms, and rearranging yields

$$(\widetilde{\gamma} - \gamma)\|s^{\widetilde{\gamma}}\|_\phi^2 \leq (\widetilde{\gamma} - \gamma)\|s^\gamma\|_\phi^2,$$

such that $\|s^\gamma\|_\phi^2$ is monotonically decreasing in $\gamma$. Moreover, it follows that

$$\gamma\big(\|s^\gamma\|_\phi^2 - \|s^{\widetilde{\gamma}}\|_\phi^2\big) \leq \frac{1}{n}\sum_{i=1}^n w_i\big(s^{\widetilde{\gamma}}(x_i) - \hat{f}^{(n)}(x_i)\big)^2 - \frac{1}{n}\sum_{i=1}^n w_i\big(s^\gamma(x_i) - \hat{f}^{(n)}(x_i)\big)^2, \tag{26}$$

showing that $\frac{1}{n}\sum_{i=1}^n w_i(s^\gamma(x_i) - \hat{f}^{(n)}(x_i))^2$ is monotonically increasing in $\gamma$.

To establish the strict monotonicity of both functions in case $(\hat{f}^{(n)}(x_1),\dots,\hat{f}^{(n)}(x_n))^T \notin \mathcal{R}(P)$, we start by showing that the minimisers $s^\gamma$ and $s^{\widetilde{\gamma}}$ cannot be identical for $0<\gamma<\widetilde{\gamma}$. To this end, assume $s^\gamma \equiv s^{\widetilde{\gamma}}$ and observe that the linear system (20) provides

$$s^\gamma(x_i)-\hat{f}(x_i) = -n\gamma w_i^{-1}\lambda_i^\gamma \quad \text{and} \quad s^\gamma(x_i)-\hat{f}(x_i) = -n\widetilde{\gamma} w_i^{-1}\lambda_i^\gamma, \tag{27}$$

$i=1,\dots,n$, where $\lambda_i^\gamma$ denotes the $i$-th coefficient of $s^\gamma$. The latter in turn yields

$$n(\widetilde{\gamma}-\gamma)w_i^{-1}\lambda_i^\gamma = 0,$$

and therefore $\lambda_i^\gamma = 0$ for $i = 1\dots,n$. This, however, implies that $s^\gamma \in \mathcal{P}_m^d$, such that the function values $\hat{f}^{(n)}(x_1),\dots,\hat{f}^{(n)}(x_n)$ in (27) are interpolated by a polynomial from the linear space $\mathcal{P}_m^d$, which contradicts the assumption $(\hat{f}^{(n)}(x_1),\dots,\hat{f}^{(n)}(x_n))^T \notin \mathcal{R}(P)$. Since $s^\gamma \neq s^{\widetilde{\gamma}}$ and the solution of (19) is unique according to Theorem 1, we even have

$$\gamma\|s^\gamma\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i\big(s^\gamma(x_i)-\hat{f}^{(n)}(x_i)\big)^2 < \gamma\|s^{\widetilde{\gamma}}\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i\big(s^{\widetilde{\gamma}}(x_i)-\hat{f}^{(n)}(x_i)\big)^2,$$

i.e. $<$ holds instead of $\leq$. Adding both inequalities and rearranging as before immediately yields

$$(\widetilde{\gamma}-\gamma)\|s^{\widetilde{\gamma}}\|_\phi^2 < (\widetilde{\gamma}-\gamma)\|s^\gamma\|_\phi^2,$$

and thus the strict monotonicity of $\|s^\gamma\|_\phi$. Finally, the strict monotonicity of $\frac{1}{n}\sum_{i=1}^n w_i(s^\gamma(x_i)-\hat{f}^{(n)}(x_i))^2$ follows by (26). □

*Choosing the Regularisation Parameter*

We will see later that to show convergence of the method it suffices to choose the regularisation parameter $\gamma$ in each step in such a way that the corresponding sequence $\{\gamma_n\}$ converges to zero quickly enough, in particular if $\gamma_n = \mathbf{o}(1/n)$. This can easily be achieved by choosing $\{\gamma_n\}$ to be an appropriate exogeneous sequence, e. g. $\gamma_n = 1/n^{1+\delta}$ for some $\delta > 0$.

However, depending on the noise model, much can be gained by choosing $\gamma_n$ adaptively to control the bumpiness of the approximant $s_n^{\gamma_n}$. Proposition 1 provides a corresponding framework: the parameter $\gamma_n$ can be identified uniquely under the weak assumption that $\widehat{F} \notin \mathcal{R}(P)$ in the following way. Finding the smoothest approximant $s_n^{\gamma_n}$ such that it deviates at the considered points $x_i$ from the noisy function values $\hat{f}^{(n)}(x_i)$ by at most $\epsilon_i$ can be stated as the *auxiliary problem*

$$\begin{aligned} &\max_{\gamma\geq 0} \quad \gamma \\ &\text{s.t.} \qquad \big|s_n^\gamma(x_i)-\hat{f}^{(n)}(x_i)\big| \leq \epsilon_i, \quad i=1,\dots,n. \end{aligned} \tag{28}$$

Problem (28) consists of a linear objective function in one dimension, which is subject to $n$ nonlinear inequality constraints. Since $s_n^\gamma$ converges to the interpolant of $\hat{f}^{(n)}(x_1),\dots,\hat{f}^{(n)}(x_n)$ for $\gamma\to 0$, as can be read off from the regularised system (20), a feasible solution to problem (28) exists. However, unlike the sum-of-squares function, the individual constraints are potentially non-monotonic in $\gamma$, and each evaluation of the constraints requires to solve the linear system (20).

This renders the problem difficult to solve and unnecessarily time-consuming if a solution is sought that is as exact as possible. Thus, as $\gamma_n$ is readjusted in each iteration upon the addition of a new point, searching for an approximate solution is sufficient. Preliminary numerical experiments indicate that appropriate values of $\gamma_n$ can be obtained by an efficient backtracking strategy, which starts with a large enough $\gamma_n$ and successively decreases this value until all constraints of (28) are met for the first time.

### 4.2 Determination of the Next Evaluation Point

To determine the next point of evaluation $x_{n+1}$, we continue similar to Jones' technique and assume that a noise-free target value $f_n^*$ has been chosen. Let $\gamma_n > 0$ be chosen appropriately. Then, let $x_{n+1}$ be the point $y \in \mathcal{X}\backslash\{x_1, \dots, x_n\}$ such that the augmented approximant $s_y^{\gamma_n} \in \mathcal{A}_{\phi,m}(\mathcal{X})$ minimises the regularised least-squares criterion to previous sample points and interpolates $f_n^*$ at the new $y$. In formal terms, we thus require that $s_y^{\gamma_n}$ solves

$$\min_{s \in \mathcal{A}_{\phi,m}(\mathcal{X})} \gamma_n \|s\|_\phi^2 + \frac{1}{n} \sum_{i=1}^{n} w_i \big(s(x_i) - \hat{f}^{(n)}(x_i)\big)^2 \qquad \text{s.t.} \quad s(y) = f_n^*, \tag{29}$$

which is a strictly convex optimisation problem on $\mathcal{A}_{\phi,m}(\mathcal{X})$ and thus admits a unique solution, cf. Theorem 1.

To simplify problem (29) in terms of the sought new point $y \in \mathcal{X}\backslash\{x_1, \dots, x_n\}$, we first rewrite the augmented approximant $s_y^{\gamma_n}$ according to

$$s_y^{\gamma_n}(x) = s_n^{\gamma_n}(x) + \big[f_n^* - s_n^{\gamma_n}(y)\big]\, l_n^{\gamma_n}(y, x), \quad x \in \mathbb{R}^d, \tag{30}$$

where $l_n^{\gamma_n}(y, \cdot) \in \mathcal{A}_{\phi,m}(\mathcal{X})$ is the radial basis function approximant that solves the constrained regularised least-squares problem

$$\min_{l(y,\cdot) \in \mathcal{A}_{\phi,m}(\mathcal{X})} \gamma_n \|l(y, \cdot)\|_\phi^2 + \frac{1}{n} \sum_{i=1}^{n} w_i \big(l(y, x_i)\big)^2 \qquad \text{s.t.} \quad l(y, y) = 1. \tag{31}$$

Representation (30) is valid since for any $y \in \mathcal{X}\backslash\{x_1, \dots, x_n\}$ both $s_n^{\gamma_n}$ and $l_n^{\gamma_n}(y, \cdot)$ are uniquely defined as solutions to the problems (19) and (31), respectively. Hence, the right-hand side of (30) is a unique well-defined element of $\mathcal{A}_{\phi,m}(\mathcal{X})$, and also satisfies the interpolation constraint in problem (29). Moreover, similar to Theorem 1, it can be shown that the approximating function $l_n^{\gamma_n}(y, \cdot)$ has the form

$$l_n^{\gamma_n}(y, x) = \sum_{i=1}^{n} \alpha_i(y)\phi(\|x - x_i\|_2) + \beta(y)\phi(\|x - y\|_2) + \sum_{j=1}^{\widetilde{m}} b_j(y) p_j(x), \quad x \in \mathbb{R}^d,$$

where the coefficients[1] $\alpha(y) = (\alpha_1(y), \dots, \alpha_n(y))^\top \in \mathbb{R}^n$, $\beta(y) \in \mathbb{R}$ and $b(y) = (b_1(y), \dots, b_{\widetilde{m}}(y))^\top \in \mathbb{R}^{\widetilde{m}}$ are defined by the linear system

$$\begin{pmatrix} \Phi + n\gamma_n W^{-1} & u_n(y) & P \\ u_n(y)^\top & \phi(0) & \pi(y)^\top \\ P^\top & \pi(y) & 0 \end{pmatrix} \begin{pmatrix} \alpha(y) \\ \beta(y) \\ b(y) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \tag{32}$$

[1] Note that these coefficients $\alpha_i(y), \beta(y), b_j(y)$ depend also on the parameter $\gamma_n$. To keep the notation simple, we will denote the dependency on $\gamma_n$ only for the approximant $s_n^{\gamma_n}$ and for the function $l_n^{\gamma_n}(y, \cdot)$.

for the matrices $\Phi \in \mathbb{R}^{n\times n}$, $P \in \mathbb{R}^{n\times \widetilde{m}}$ and $W \in \mathbb{R}^{n\times n}$ introduced before, and the corresponding vectors $u_n(y) = (\phi(\|x_1 - y\|_2), \dots, \phi(\|x_n - y\|_2))^\top \in \mathbb{R}^n$ and $\pi(y) = (p_1(y), \dots, p_{\widetilde{m}}(y))^\top \in \mathbb{R}^{\widetilde{m}}$.

By inserting representation (30) into the objective function of (29), the latter can then be reformulated as

$$\begin{aligned}
&\gamma_n \|s_y^{\gamma_n}\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i \big(s_y^{\gamma_n}(x_i) - \hat{f}^{(n)}(x_i)\big)^2 \\
&\quad = \gamma_n \Big( \|s_n^{\gamma_n}\|_\phi^2 + 2\big[f_n^* - s_n^{\gamma_n}(y)\big]\langle s_n^{\gamma_n}, l_n^{\gamma_n}(y,\cdot)\rangle_\phi + \big[f_n^* - s_n^{\gamma_n}(y)\big]^2 \|l_n^{\gamma_n}(y,\cdot)\|_\phi^2 \Big) \\
&\qquad + \frac{1}{n}\sum_{i=1}^n w_i \Big( \big(s_n^{\gamma_n}(x_i) - \hat{f}^{(n)}(x_i)\big)^2 + \big[f_n^* - s_n^{\gamma_n}(y)\big]^2 \big(l_n^{\gamma_n}(y,x_i)\big)^2 \\
&\qquad + 2\big(s_n^{\gamma_n}(x_i) - \hat{f}^{(n)}(x_i)\big)\big[f_n^* - s_n^{\gamma_n}(y)\big] l_n^{\gamma_n}(y,x_i) \Big) \\
&\quad = \gamma_n \|s_n^{\gamma_n}\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i \big(s_n^{\gamma_n}(x_i) - \hat{f}^{(n)}(x_i)\big)^2 \\
&\qquad + \big[f_n^* - s_n^{\gamma_n}(y)\big]^2 \Big( \gamma_n \| l_n^{\gamma_n}(y,\cdot)\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i \big(l_n^{\gamma_n}(y,x_i)\big)^2 \Big),
\end{aligned} \tag{33}$$

where the last equation holds by definition of the semi-inner product (6) and the relation $s_n^{\gamma_n}(x_i) - \hat{f}^{(n)}(x_i) = -n\gamma_n w_i^{-1}\lambda_i$, $i = 1, \dots, n$, due to (20), which both together yield

$$\begin{aligned}
\langle s_n^{\gamma_n}, l_n^{\gamma_n}(y,\cdot)\rangle_\phi &= \sum_{i=1}^n \lambda_i l_n^{\gamma_n}(y,x_i) \\
&= -\frac{1}{n\gamma_n}\sum_{i=1}^n w_i \big(s_n^{\gamma_n}(x_i) - \hat{f}^{(n)}(x_i)\big) l_n^{\gamma_n}(y,x_i).
\end{aligned}$$

Now, the first two terms on the right-hand side of equation (33) are independent of $y$ and correspond to the objective function for constructing the approximant $s_n^{\gamma_n}$, cf. problem (19). To find the new point $y$, it thus suffices to consider the last term in (33). However, by the semi-inner product (6) and the linear system (32), implying $l_n^{\gamma_n}(y,x_i) = -n\gamma_n w_i^{-1}\alpha_i(y)$ for $i = 1, \dots, n$, it holds

$$\begin{aligned}
&\gamma_n \|l_n^{\gamma_n}(y,\cdot)\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i \big(l_n^{\gamma_n}(y,x_i)\big)^2 \\
&\quad = \gamma_n \left( \sum_{i=1}^n \alpha_i(y) l_n^{\gamma_n}(y,x_i) + \beta(y) l_n^{\gamma_n}(y,y) + \frac{1}{n}\sum_{i=1}^n w_i \big(l_n^{\gamma_n}(y,x_i)\big)^2 \right) \\
&\quad = \gamma_n \beta(y).
\end{aligned}$$

Therefore, we can conclude that solving the required problem (29) is equivalent to minimising the nonnegative function

$$g_n^{\gamma_n}(y) := \mu_n^{\gamma_n}(y)\big[f_n^* - s_n^{\gamma_n}(y)\big]^2, \qquad y \in \mathcal{X}\backslash\{x_1, \dots, x_n\} \tag{34}$$

with respect to $y$, where the function $\mu_n^{\gamma_n} : \mathcal{X}\backslash\{x_1,\dots,x_n\} \to \mathbb{R}$ is defined for $\gamma_n > 0$ by

$$\mu_n^{\gamma_n}(y) := \|l_n^{\gamma_n}(y,\cdot)\|_\phi^2 + \frac{1}{n\gamma_n}\sum_{i=1}^n w_i\big(l_n^{\gamma_n}(y,x_i)\big)^2 = \beta(y). \tag{35}$$

Note the resemblance of the functions $g_n^{\gamma_n}$ and $\mu_n^{\gamma_n}$ to their deterministic counterparts (14) and (15), respectively. In particular, since $l_n^{\gamma_n}(y,\cdot)$ is well-defined for $y \in \mathcal{X}\backslash\{x_1,\dots,x_n\}$, so are both functions $g_n^{\gamma_n}$ and $\mu_n^{\gamma_n}$.

In an analogous manner to Gutmann (2001a), Section 5 (cf. also Gutmann (2001c), Proposition 4.12), the function $\mu_n^{\gamma_n}$ (and thus $g_n^{\gamma_n}$) can be equivalently expressed as follows, which allows for a more intuitive interpretation as well as a more efficient computation.

**Proposition 2** *For $\gamma_n > 0$, the function $v_n^{\gamma_n}$ defined by*

$$v_n^{\gamma_n}(y) := \left[\phi(0) - \begin{pmatrix} u_n(y) \\ \pi(y) \end{pmatrix}^\top \begin{pmatrix} \Phi + n\gamma_n W^{-1} & P \\ P^\top & 0 \end{pmatrix}^{-1} \begin{pmatrix} u_n(y) \\ \pi(y) \end{pmatrix}\right], \qquad y \in \mathbb{R}^d,$$

*is identical to $1/\mu_n^{\gamma_n}$ on $y \in \mathbb{R}^d\backslash\{x_1,\dots,x_n\}$. Moreover, $v_n^{\gamma_n}$ can be continuously extended at the sample points $x_1,\dots,x_n$ by the finite values $v_n^{\gamma_n}(x_i) = 1/\mu_n^{\gamma_n}(x_i)$, $i = 1,\dots,n$.*

*Proof* For any $y \in \mathbb{R}^d\backslash\{x_1,\dots,x_n\}$, the proof follows in a straightforward manner by using the definition of $\mu_n(y)$ in (35) and solving the equations in the linear system (32) for the coefficient $\beta(y)$ by rearranging and applying the Schur complement of the invertible block $A_n^{\gamma_n}$ in the matrix $A_n^{\gamma_n}(y)$. □

Since the function $v_n^{\gamma_n}$ is positive and finite on $\mathcal{X}$ for $\gamma_n > 0$, it can be understood as a measure of uncertainty in the approximating model $s_n^{\gamma_n}$ to $f$. In particular, the error $v_n^{\gamma_n}(y)$ at any $y$ is influenced by the distance to the sample points $x_1,\dots,x_n$ as well as the inherent noise resulting from inexact function values, which is most notably reflected by the fact that $v_n^{\gamma_n}(x_i) > 0$ for $i = 1,\dots,n$.

*Remark 3* By the same argument as given in Remark 1, definition (35) implies that the function $\mu_n^{\gamma_n}$ is positive on $\mathcal{X}\backslash\{x_1,\dots,x_n\}$.

Moreover, by definition (35) and Cramer's rule to solve the linear system (32), we have

$$\mu_n^{\gamma_n}(y) = \frac{\det A_n^{\gamma_n}}{\det A_n^{\gamma_n}(y)}, \qquad y \in \mathcal{X}\backslash\{x_1,\dots,x_n\},$$

where $A_n^{\gamma_n}$ and $A_n^{\gamma_n}(y)$ denote the nonsingular matrices on the left-hand sides of the linear systems (20) and (32), respectively, that is

$$A_n^{\gamma_n} := \begin{pmatrix} \Phi + n\gamma_n W^{-1} & P \\ P^\top & 0 \end{pmatrix} \quad \text{and} \quad A_n^{\gamma_n}(y) := \begin{pmatrix} \Phi + n\gamma_n W^{-1} & u_n(y) & P \\ u_n(y)^\top & \phi(0) & \pi(y)^\top \\ P^\top & \pi(y) & 0 \end{pmatrix}. \tag{36}$$

Since $\det A_n^{\gamma_n}$ is a nonzero constant and $\lim_{y\to x_i} \det A_n^{\gamma_n}(y) \neq 0$ ($i \in \{1,\dots,n\}$) by the positivity of $v_n^{\gamma_n}$, it thus holds

$$\lim_{y\to x_i} \mu_n^{\gamma_n}(y) < \infty, \qquad i = 1,\dots,n.$$

Hence, even though $\mu_n^{\gamma_n}$ is not defined at the sample points $x_1, \ldots, x_n$, it can be continuously extended at these points by the positive and finite values

$$\mu_n^{\gamma_n}(x_i) = \frac{\det A_n^{\gamma_n}}{\det A_n^{\gamma_n}(x_i)}, \qquad i = 1, \ldots, n, \tag{37}$$

due to the continuity of the determinant.

Finally, based on Proposition 2, the function $h_n^{\gamma_n}$ defined by

$$h_n^{\gamma_n}(y) := \frac{v_n^{\gamma_n}(y)}{\left[s_n^{\gamma_n}(y) - f_n^*\right]^2}, \qquad y \in \mathbb{R}^d,$$

can be shown to be identical to $1/g_n^{\gamma_n}$ on $y \in \mathbb{R}^d \backslash \{x_1, \ldots, x_n\}$, with continuously extended values $h_n^{\gamma_n}(x_i) = 1/g_n^{\gamma_n}(x_i)$, $i = 1, \ldots, n$.

### 4.3 Choice of Target Value

As the target value $f_n^*$ has the same functionality as in Algorithm 1, its choice determines the location of the next evaluation point $x_{n+1}$, minimising $g_n^{\gamma_n}$ on $\mathcal{X}\backslash\{x_1, \ldots, x_n\}$ for fixed $\gamma_n > 0$, in a similar fashion to Algorithm 1. For $f_n^* < \min_{y \in \mathcal{X}} s_n^{\gamma_n}(y)$ we usually interpret $f_n^*$ as the next target function value, i. e. we want to compute an argument $x_{n+1}$ with objective function value close to $f_n^*$. Unfortunately, however, it remains unclear at this point whether a choice $f_n^* < \min_{y \in \mathcal{X}} s_n^{\gamma_n}(y)$ is also sufficient to guarantee that a global minimiser of $g_n^{\gamma_n}$ on $\mathcal{X}\backslash\{x_1, \ldots, x_n\}$ exists, i.e. that a global minimiser of $g_n^{\gamma_n}$ over $\mathcal{X}$ does not coincide with any of the sample points $x_1, \ldots, x_n$, as in the case of interpolation, or whether a further condition is required. The main issue here is due to the fact that $g_n^{\gamma_n}$ is continuously extendable at the points $x_i$, $i = 1, \ldots, n$, by a finite value, cf. equation (37), which then implies that $g_n^{\gamma_n}(y)$ does not tend to infinity anymore as $y$ approaches any $x_i$.

In any case, though, note that for an admissible choice $f_n^* < \min_{y \in \mathcal{X}} s_n^{\gamma_n}(y)$, we may draw the same conclusions for $g_n^{\gamma_n}$ as for $g_n$ in that, for $f_n^* = -\infty$, the minimisation of $g_n^{\gamma_n}$ on $\mathcal{X}\backslash\{x_1, \ldots, x_n\}$ reduces to minimising $\mu_n^{\gamma_n}$ on $\mathcal{X}\backslash\{x_1, \ldots, x_n\}$. Moreover, due to the identities $v_n^{\gamma_n} = 1/\mu_n^{\gamma_n}$ and $h_n^{\gamma_n} = 1/g_n^{\gamma_n}$, as established in the previous subsection, the minimisers of $\mu_n^{\gamma_n}$ and $g_n^{\gamma_n}$ correspond to the maximisers of $v_n^{\gamma_n}$ and $h_n^{\gamma_n}$ on $\mathcal{X}\backslash\{x_1, \ldots, x_n\}$, respectively. Hence, if $-\infty < f_n^* < \min_{y \in \mathcal{X}} s_n^{\gamma_n}(y)$, we can equivalently maximise the utility function $h_n^{\gamma_n}$ and, if $f_n^* = -\infty$, the respective function $v_n^{\gamma_n}$ on $\mathcal{X}\backslash\{x_1, \ldots, x_n\}$.

In case we end up in a situation in which one of the $x_i$ as a global minimiser of $g_n^{\gamma_n}$, we would revisit $x_{n+1} = x_i$ instead of finding a new point. In this case, there are various ways ahead, bearing in mind that we want to construct a sequence $(x_n)_n$ that is dense in $\mathcal{X}$. We could thus consider e. g. a Delaunay triangulation induced by the $x_1, \ldots, x_n$ and then choose $x_{n+1}$ as the center point of the largest simplex in the triangulation, or update $f_n^*$ to a smaller value and repeat the iteration. Since this is not of importance in the convergence proofs and to keep the exposition succinct, we will not describe how to handle this case when stating our algorithm in the next subsection.

### 4.4 Summary of the RBF Method for Noisy Objective Functions

In summary, the RBF method for minimising a noisy objective function $\hat{f} : \mathcal{X} \to \mathbb{R}$ on a compact set $\mathcal{X}$ can be formulated by the following algorithm.

**Algorithm 2** *(RBF Method for Noisy Objective Functions).*

*0.* ***Initial step:***
- *Choose a conditionally positive definite radial basis function $\phi$ of order $m$.*
- *Generate a $\mathcal{P}_m^d$-unisolvent set of points $\{x_1, \dots, x_{n_0}\} \subset \mathcal{X}$.*
- *Evaluate $\hat{f}^{(n_0)}$ at the points $x_1, \dots, x_{n_0}$, and set $n = n_0$.*
- *Choose $x^{(n_0)} \in \arg\min_{x_i} \hat{f}^{(n_0)}(x_i)$.*

*1.* ***Iteration step:***

***while*** $n \leq n^{\max}$ ***do***
- *Choose $\gamma_n > 0$.*
- *Construct the approximant $s_n^{\gamma_n} \in \mathcal{A}_{\phi,m}(\mathcal{X})$ solving*

$$\min_{s \in \mathcal{A}_{\phi,m}(\mathcal{X})} \gamma_n \|s\|_\phi^2 + \frac{1}{n} \sum_{i=1}^{n} w_i \big(s(x_i) - \hat{f}^{(n)}(x_i)\big)^2.$$

- *Choose an admissible target value $f_n^* \in \big[-\infty, \min_{y \in \mathcal{X}} s_n^{\gamma_n}(y)\big]$.*
- *Determine $x_{n+1}$, which is the value of $y$ solving*

$$\min_{y \in \mathcal{X} \setminus \{x_1, \dots, x_n\}} \mu_n^{\gamma_n}(y) \big[f_n^* - s_n^{\gamma_n}(y)\big]^2.$$

- *Evaluate $\hat{f}^{(n)}$ at $x_{n+1}$, resulting in $\hat{f}^{(n)}(x_{n+1})$.*
- *For $i = 1, \dots, n+1$, update the evaluation $\hat{f}^{(n)}(x_i)$ at $x_i$ to $\hat{f}^{(n+1)}(x_i)$.*
- *Set $x^{(n+1)}$ as a point $z$ from the set $\{x_1, \dots, x_{n+1}\}$ with $\hat{f}^{(n+1)}(z) \leq \hat{f}^{(n+1)}(x_i)$, $i = 1, \dots, n+1$.*
- *Set $n = n + 1$.*

***end while***

## 5 Convergence of Method

As Gutmann's original method, our RBF method for noisy objective functions is a purely deterministic sequential sampling algorithm. For a given set of noisy function values, the construction of an approximant and the subsequent selection of a new evaluation point is carried out independently of any source of randomness. To show convergence of the method to the global minimum of any continuous function $f$ by means of noisy objective function values $\hat{f}^{(n)}(x_i)$, our main task is thus to establish the density of the sequence of generated iterates $\{x_n\}$ in $\mathcal{X}$, cf. Törn and Žilinskas (1989), Theorem 1.3. We can therefore state the following obvious theorem for the convergence of Algorithm 2.

**Theorem 2** *Let $f$ be a continuous function on the compact set $\mathcal{X}$ with minimum function value $f^*$. Suppose $\epsilon_n \to 0$ for $n \to \infty$ in the case of fixed noise and $\max_{1\leq i\leq n} \epsilon_i^{(n)} \to 0$ for $n \to \infty$ in case of vanishing iterative noise. Then, Algorithm 2 provides a sequence $\{x^{(n)}\}$ with $\lim_{n\to\infty} \hat{f}^{(n)}(x^{(n)}) = \lim_{n\to\infty} f(x^{(n)}) = f^*$ if the algorithm generates a sequence of points $\{x_n\}$ that is dense in $\mathcal{X}$.*

As it turns out in the below, a key result that allows to establish the density of the sequence of generated points $\{x_n\}$ is the relationship

$$\lim_{n\to\infty} n\gamma_n = 0,$$

i.e. $\gamma_n = \mathbf{o}(1/n)$. To this end, we first address the influence of the error bounds $\{\epsilon_n\}$ on the sequence $\{n\gamma_n\}$, provided that $\gamma_n$ is chosen in each iteration according to the auxiliary problem (28). We then present several convergence results, followed by a proof of convergence of the main statement.

### 5.1 Assumptions on Error Bounds

One possibility for establishing density of the iterates $x_1, \ldots, x_n$ is to resort to the available convergence results of Gutmann's method and show that these pertain if the exact function values $f(x_i)$ are replaced by the noisy observations $\hat{f}^{(n)}(x_i)$ $(1 \leq i \leq n)$. An indispensable assumption is thus that the involved level of noise decreases to zero over the course of the optimisation. As one may already conjecture from the construction of regularised least-squares approximants through (20), this will be required to adopt Gutmann's proof of convergence for noisy function values. Nevertheless, in the vanishing noise model we have already that for a natural choice of $\gamma_n$, the sequence $\{n\gamma_n\}$ converges to zero if we require that $\epsilon_i^{(n)} \to 0$ fast enough as $n, i \to \infty$, as the following theorem shows.

**Theorem 3** *Let $\phi$ be a conditionally positive definite radial basis function spline of order $m$, and let $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ be a $\mathcal{P}_m^d$-unisolvent set. Let $s_n^{\gamma_n} \in \mathcal{A}_{\phi,m}(\mathcal{X})$ denote the unique optimal solution of the regularised least-squares problem* (19)*, where the regularisation parameter $\gamma_n > 0$ solves*

$$\begin{aligned} \max_{\substack{\gamma>0 \\ s_n^\gamma \in \mathcal{A}_{\phi,m}}} \quad & \gamma \\ s.t. \quad & \left|s_n^\gamma(x_i) - \hat{f}^{(n)}(x_i)\right| \leq \epsilon_i^{(n)}, \quad i = 1, \ldots, n, \end{aligned} \tag{38}$$

*for some positive error bounds $\epsilon_i^{(n)}$.*

1. *Let the function values $\hat{f}^{(n)}(x_1), \ldots, \hat{f}^{(n)}(x_n)$ be such that $\widehat{F}_n = (\hat{f}^{(n)}(x_1), \ldots, \hat{f}^{(n)}(x_n))^\top \notin \mathcal{R}(P)$ for all $n$ for the corresponding polynomial basis matrix $P$. Further, assume that $\max_{1\leq i\leq n} \epsilon_i^{(n)} \to 0$ for $n \to \infty$ and that the sequence $\{S_n\}$ with*

$$S_n := \inf_{\substack{s\in\mathcal{A}_{\phi,m} \\ \|s\|_\phi=0}} \sum_{i=1}^n w_i(s(x_i) - \hat{f}^{(n)}(x_i))^2$$

*is bounded. Then, $\gamma_n = O(1/n)$.*

2. *Assume that* $\sum_{i=1}^{n}(w_i\epsilon_i^{(n)})^2 = \boldsymbol{o}(1/n)$ *and that the sequence* $\{\gamma_n\}$ *is non-increasing. Then* $\gamma_n = o(1/n)$.

*Proof* 1. For fixed $n \in \mathbb{N}$, note that problem (19) with $\gamma > 0$ constitutes a scalarisation of the bi-objective optimisation problem

$$\min_{s\in\mathcal{A}_{\phi,m}(\mathcal{X})} \Big(\|s\|_\phi^2\,, \sum_{i=1}^{n} w_i\big(s(x_i) - \hat{f}^{(n)}(x_i)\big)^2\Big), \tag{39}$$

with corresponding weight vector $(\gamma, 1/n)^\top$. Consequently, since $\gamma > 0$, any optimal solution $s^\gamma$ of the scalarised problem (19) is Pareto optimal for problem (39), and since $\mathcal{A}_{\phi,m}(\mathcal{X})$ is a convex set and both objective functions in (39) are convex in $s^\gamma$, there is some positive weight vector for any Pareto optimal point of (39) such that it is an optimal solution of the scalarisation (19), see, e.g., Ehrgott (2005), Theorem 4.1.
Since for any $\gamma_n > 0$, the optimal solution $s_n^{\gamma_n}$ of (19) is unique, the mapping

$$G_n : \gamma_n \mapsto \Big(\|s_n^{\gamma_n}\|_\phi^2\,, \sum_{i=1}^{n} w_i\big(s_n^{\gamma_n}(x_i) - \hat{f}^{(n)}(x_i)\big)^2\Big),$$

assigning $\gamma_n$ to a point of the set of Pareto optimal solutions, is well-defined. Due to Proposition 1, the functions $\|s_n^{\gamma_n}\|_\phi^2$ and $\sum_{i=1}^{n} w_i(s_n^{\gamma_n}(x_i) - \hat{f}^{(n)}(x_i))^2$ are also continuous and strictly monotone in $\gamma_n$, such that each function value of $G_n$ is attained uniquely by changing $\gamma_n$. In particular, by increasing $\gamma_n$, we strictly decrease $\|s_n^{\gamma_n}\|_\phi^2$ and strictly increase $\sum_{i=1}^{n} w_i(s_n^{\gamma_n}(x_i) - \hat{f}^{(n)}(x_i))^2$, and vice versa, where the extreme points of $G_n$ are given for $\gamma_n \to 0$ by the interpolant to $\hat{f}^{(n)}(x_1), \dots, \hat{f}^{(n)}(x_n)$, and for $\gamma_n \to \infty$ by a Pareto optimal solution $s_n^\infty$ with $\|s_n^\infty\|_\phi^2 = 0$ that minimizes the term $\sum_{i=1}^{n}(s_n(x_i) - f^{(n)}(x_i))^2$ over all $s$ with $\|s\|_\phi = 0$. In particular, we can define $G_n(\infty) = (0, S_n)$ and note that $S_n \le C$ for some $C > 0$.
Thus, the graph of $G_n$ can be identified with the graph of a function $\varphi_n$ with images in $\mathbb{R}_{\ge 0}$, where for a given $\xi_n = \|s_n^{\gamma_n}\|_\phi^2$ with unique $\gamma_n = \gamma_n(\xi_n)$, we set $\varphi(\xi_n) = \sum_{i=1}^{n} w_i(s_n^{\gamma_n}(x_i) - \hat{f}^{(n)}(x_i))^2$. Similarly, for given $\gamma_n$, we let $\xi_n(\gamma_n)$ be the unique $\xi_n$ with $\gamma_n = \gamma_n(\xi_n(\gamma_n))$. By definition, both functions $\varphi_n(\xi_n)$ and $\xi_n(\gamma_n)$ are thus continuous, strictly convex, monotonically decreasing, and thus differentiable almost everywhere.
Since, by assumption, $\max_{1\le i\le n} \epsilon_i^{(n)} \to 0$ as $n \to \infty$, it follows $\varphi_n(\xi_n) \to 0$, Moreover, a subderivative of $\varphi_n$ at $\xi_n$ is given by $-n\gamma_n$. Due to the uniqueness of the weight vector $(\gamma_n, 1/n)$, the subderivative is unique, and thus $\varphi_n$ is differentiable at $\xi_n$ with $\varphi_n'(\xi_n) = -n\gamma_n$.
Suppose now that the sequence $\{\gamma_n n\}$ is unbounded. By the subgradient inequality, this precludes $\varphi_n(0) = S_n \le C$ for infinitely many $n$, unless $\xi_n$ converges to 0, However, the $\xi_n$ are monotonically increasing as the $\gamma_n$ are monotonically decreasing.

2. For given $n$ and $\gamma > 0$, let $\lambda = \lambda^{(n,\gamma)}$ and $c = c^{(n,\gamma)}$ denote the unique solution of the linear system (20), i.e. $\lambda$ and $c$ are the coefficients of the optimal solution $s_n^\gamma$ of (19). Then, considering the first block of equation (20), we especially obtain

$$\Phi\lambda + n\gamma W^{-1}\lambda + Pc = \hat{F}.$$

The $i$-th row of this equation shows

$$s_n^\gamma(x_i) + n\gamma w_i^{-1}\lambda_i = \hat{f}^{(n)}(x_i)$$

and thus

$$|s_n^\gamma(x_i) - \hat{f}^{(n)}(x_i)| = |n\gamma w_i^{-1}\lambda_i|.$$

Therefore, the $i$-th constraint in (28) is equivalent to

$$|n\gamma\lambda_i| \leq w_i\epsilon_i^{(n)}.$$

Let $\epsilon^{(n)} := (\epsilon_1^{(n)}, \ldots, \epsilon_n^{(n)})^\top$, Since $\Phi$ only contains non-negative entries and is conditionally positive definite, we arrive at

$$n^2\gamma^2\lambda^\top\Phi\lambda \leq {\epsilon^{(n)}}^\top W\Phi W\epsilon^{(n)}$$

for all $\gamma$ feasible for (38). Since we assume that the optimal solutions $\gamma_n$ provide a non-increasing sequence, Proposition 1c) yields for all $n \geq n_0$ that

$$||s_n^{\gamma_n}||_\phi^2 \geq ||s_{n_0}^{\gamma_{n_0}}||_\phi^2.$$

Taking into account that

$$\lambda^\top\Phi\lambda = ||s_n^{\gamma_n}||_\phi^2$$

we get for all $n \geq n_0$ that

$$\lambda^\top\Phi\lambda = ||s_n^{\gamma_n}||_\phi^2 \geq ||s_{n_0}^{\gamma_{n_0}}||_\phi^2.$$

Combining the above inequalities, we obtain:

$$n^2\gamma_n^2||s_{n_0}^{\gamma_{n_0}}||_\phi^2 \leq n^2\gamma_n^2\lambda^\top\Phi\lambda \leq {\epsilon^{(n)}}^\top W\Phi W\epsilon^{(n)}.$$

We continue by bounding the Frobenius norm of $\Phi$ by

$$||\Phi||_F^2 = \sum_{i,j}\Phi_{i,j}^2 \leq n^2 \max_{u,v\in X}\phi(||u-v||)^2 =: n^2\bar{\phi}^2$$

and thus

$${\epsilon^{(n)}}^\top W\Phi W\epsilon^{(n)} \leq n\bar{\phi}||W\epsilon^{(n)}||_2^2 = n\bar{\phi}\sum_i(w_i\epsilon_i^{(n)})^2.$$

Summarizing we finally obtain the inequality

$$n^2\gamma_n^2 \leq n\frac{\bar{\phi}}{||s_{n_0}^{\gamma_{n_0}}||_\phi^2}\sum_i(w_i\epsilon_i^{(n)})^2,$$

and thus

$$\gamma_n^2 \leq \frac{\bar{\phi}}{||s_{n_0}^{\gamma_{n_0}}||_\phi^2}\frac{1}{n}\sum_i(w_i\epsilon_i^{(n)})^2.$$

From this, we immediately see that for $\gamma_n = \mathbf{o}(1/n)$ it is sufficient to require $\frac{1}{n}\sum_i(w_i\epsilon_i^{(n)})^2 = \mathbf{o}(1/n^2)$, i.e. $\sum_i(w_i\epsilon_i^{(n)})^2 = \mathbf{o}(1/n)$. □

### 5.2 Convergence Results

Besides assuming $n\gamma_n \to 0$ as $n \to \infty$, we further require the target values $f_n^*$ to be set sufficiently low compared to the approximating surfaces $s_n^{\gamma_n}$ in order to achieve convergence of the RBF method for noisy objective functions, cf. Gutmann (2001a), condition (4.2) (or Gutmann (2001c), condition (4.16))[2]. Due to the presence of noise, the critical thresholds for $f_n^*$ need to be adjusted marginally to guarantee convergence of the method in a similar fashion as Gutmann. To this end, we let, for infinitely many $n \in \mathbb{N}$, the target values $f_n^*$ satisfy

$$f_n^* < \min_{y \in \mathcal{X}} \Big[ s_n^{\gamma_n}(y) - \tau \|s_n^{\gamma_n}\|_\infty \big[ \Delta_n(y) + \widetilde{w}_n^{-1/2}(y) \big]^{\rho/2} \Big], \tag{40}$$

where, as in the noise-free counterpart, $\tau > 0$ and $\rho \geq 0$ are constants with $\rho < 1$, for $\phi(r) = r$, and $\rho < 2$, otherwise, and $\Delta_n$ denotes the minimum distance function

$$\Delta_n(y) := \min_{1 \leq i \leq n} \|y - x_i\|_2, \quad y \in \mathcal{X}. \tag{41}$$

For given $y \in \mathcal{X}$, the function $\widetilde{w}_n(y)$ gives the weight $w_i$ of the sample point $x_i$ that is closest to $y$, i.e. for $i(y) = \arg\min_{1 \leq i \leq n} \|y - x_i\|_2$, we have

$$\widetilde{w}_n(y) := w_{i(y)}, \tag{42}$$

with the convention that the largest $i(y)$ is selected among the minimising indices if arg min is not unique.

Finally, note that the convergence of the method is restricted to the choice of radial basis function $\phi$ as its proof requires to bound the sequence $\{\mu_n^{\gamma_n}(y)\}$ uniformly from above for any $y \in \mathbb{R}^d$ that is bounded away from the points in the sequence $\{x_n\}$, cf. Lemma 4. This, however, can be shown if there is a function that takes the value 1 at $y$ and zero outside a neighbourhood of $y$, and that belongs to the corresponding native space $\mathcal{N}_\phi(\mathbb{R}^d)$ of $\phi$, as defined as follows, cf. Gutmann (2001a), Definition 5 (or Gutmann (2001c), Definition 3.10).

**Definition 1** Let $\phi$ be a conditionally positive definite radial basis function of order $m$, and $\mathcal{D} \subset \mathbb{R}^d$. Then, a function $f : \mathcal{D} \to \mathbb{R}$ belongs to the native space $\mathcal{N}_\phi(\mathcal{D})$ if and only if for any $\mathcal{P}_m^d$-unisolvent set $\{x_1, \ldots, x_n\} \subset \mathcal{D}$ the optimal interpolant $s \in \mathcal{A}_{\phi,m}(\mathcal{D})$ to $f$ at these points satisfies

$$\|s\|_\phi \leq C_f,$$

where $C_f$ is a nonnegative constant that only depends on $f$.

A useful criterion for a function to be in the native space of a radial basis function can be given for surface splines by the following theorem, cf. Gutmann (2001a), Propositions 6 and 10 (Gutmann (2001c), Theorem 3.19). An extension to (inverse) multiquadric and Gaussian radial basis functions is not possible since the respective native spaces do not contain any nonzero functions with compact support, see Gutmann (2001c), Section 6.4.

[2] Note that, throughout this section we refer to both Gutmann's publication and his dissertation, as the results in the latter are sometimes formulated in a slightly more general way.

**Theorem 4** *Let $\phi$ be a conditionally positive definite surface spline of order $m$ from Table 1, and let*

$$\nu_d = \begin{cases} (d+\nu+1)/2 & \text{if } d+\nu \text{ is odd,} \\ (d+\nu)/2 & \text{if } d+\nu \text{ is even.} \end{cases}$$

*If $f \in C^{\nu_d}(\mathcal{D})$, where (i) $\mathcal{D} \subset \mathbb{R}^d$ is compact, or (ii) $\mathcal{D} = \mathbb{R}^d$ and $f$ has compact support, then $f \in \mathcal{N}_\phi(\mathcal{D})$.*

In the case of spline type radial basis functions, Gutmann's main convergence result, stating that the generated sequence is dense in $\mathcal{X}$, cf. Gutmann (2001a), Theorem 7 (or Gutmann (2001c), Theorem 4.5), can now be formulated in the noisy setup as follows. For a proof of the statement, see Subsection 5.3.

**Theorem 5** *Let $\phi$ be a conditionally positive definite surface spline of order $m$ from Table 1, and let $\{x_n\}$ be the sequence of iterates generated by Algorithm 2. Further, let $s_n^{\gamma_n}$ with $\gamma_n > 0$ be the optimal regularised least-squares approximant from $\mathcal{A}_{\phi,m}(\mathcal{X})$ to the data $(x_i, \hat{f}^{(n)}(x_i))$, $i = 1, \ldots, n$, with corresponding weights $w_i$ bounded away from zero. Assume that, for infinitely many $n \in \mathbb{N}$, the choice of $f_n^*$ satisfies (40), where $\tau$, $\Delta_n$, $\rho$ and $\widetilde{w}_n$ are given as above, and that $n\gamma_n \to 0$ as $n \to \infty$. Then, the sequence $\{x_n\}$ is dense in $\mathcal{X}$.*

In view of Gutmann (2001a), Corollary 8 (or Gutmann (2001c), Corollary 4.6), we can conclude the following particular convergence result from Theorems 2 and 5, due to the finiteness of the right-hand side in assumption (40) for any $n \in \mathbb{N}$.

**Corollary 1** *Let $\phi$ and $m$ be as in Theorem 5. Further, let $f$ be continuous with minimal function value $f^*$, and assume that, for infinitely many $n \in \mathbb{N}$, it holds $f_n^* = -\infty$. Suppose that we have iterative noise, i.e. $\left|f(x_i) - \hat{f}^{(n)}(x_i)\right| \leq \epsilon_i^{(n)}$ for all iterates $x_i$, with $\max_{1 \leq i \leq n} \epsilon_i^{(n)} \to 0$ for $n \to \infty$, and $\gamma_n = o(1/n)$ holds. Then, we have $\lim_{n\to\infty} f(x^{(n)}) = f^*$ for the sequence $\{x^{(n)}\}$ constructed by Algorithm 2.*

To derive a further convergence result applying to functions $f$ in the native space and under particular assumptions on the error bounds, we first show that for sufficiently large $n$ the maximum norm of the approximating surface can be bounded, cf. Gutmann (2001a), Lemma 9 (or Gutmann (2001c), Lemma 4.7), for the equivalent case of interpolation. The lemma below assumes that $f$ is from the corresponding native space and uses the norm $\|\cdot\|_{\mathcal{N}_\phi}$ on this native space, as introduced by Schaback (1999).

**Lemma 1** *Let $\{x_n\}$ be a sequence in $\mathcal{X}$ with pairwise different points such that $\{x_1, \ldots, x_{n_0}\}$ is $\mathcal{P}_m^d$-unisolvent. For $n \geq n_0$, let $s_n^{\gamma_n}$ with $\gamma_n > 0$ denote the optimal regularised least-squares approximant to $\hat{f}$ at $x_1, \ldots, x_n$, where the respective weights $w_1, \ldots, w_n$ are bounded away from zero. Further, let $f \in \mathcal{N}_\phi(\mathcal{X})$, and assume that $n\gamma_n \leq n_0\gamma_{n_0}$ for sufficiently large $n$. Then, for $n$ large enough,*

$$\|s_n^{\gamma_n}\|_\infty \leq \frac{1}{\sqrt{\alpha_1}} \Big( \|f\|_{\mathcal{N}_\phi}^2 + \frac{1}{n\gamma_n} \sum_{i=1}^{n} w_i \big(f(x_i) - \hat{f}^{(n)}(x_i)\big)^2 \Big)^{1/2} + \|f\|_\infty, \qquad (43)$$

*where $\alpha_1$ is a constant depending on $x_1, \ldots, x_{n_0}$.*

*Proof* Fix $n \in \mathbb{N}$, and let $y$ be any point in $\mathcal{X}\backslash\{x_1, \dots, x_n\}$. For $\gamma_n > 0$, let $\tilde{s}_n^{\gamma_n}$ be the optimal regularised least-squares approximant from $\mathcal{A}_{\phi,m}(\mathcal{X})$ to $(x_i, \hat{f}^{(n)}(x_i))$, $i = 1, \dots, n$, with corresponding weights $w_i$ bounded away from zero, and subject to $\tilde{s}_n^{\gamma_n}(y) = f(y)$. Analogous to the derivation in Subsection 4.2, the approximant can thus be rewritten as

$$\tilde{s}_n^{\gamma_n}(x) = s_n^{\gamma_n}(x) + \big[f(y) - s_n^{\gamma_n}(y)\big] l_n^{\gamma_n}(y, x), \qquad x \in \mathbb{R}^d, \tag{44}$$

where $l_n^{\gamma_n}(y, \cdot)$ is the optimal regularised least-squares approximant to $(x_i, 0)$, with respective weights $w_i$, and subject to $l_n^{\gamma_n}(y, y) = 1$. Moreover, it follows that

$$\begin{aligned}
&\gamma_n \|\tilde{s}_n^{\gamma_n}\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i \big(\tilde{s}_n^{\gamma_n}(x_i) - \hat{f}^{(n)}(x_i)\big)^2 \\
&\quad = \gamma_n \|s_n^{\gamma_n}\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i \big(s_n^{\gamma_n}(x_i) - \hat{f}^{(n)}(x_i)\big)^2 + \big[f(y) - s_n^{\gamma_n}(y)\big]^2 \gamma_n \mu_n^{\gamma_n}(y),
\end{aligned} \tag{45}$$

where the positive function $\mu_n^{\gamma_n}$ of the approximant $l_n^{\gamma_n}(y, \cdot)$ is given by (35). Equality (45) thus yields

$$\big[f(y) - s_n^{\gamma_n}(y)\big]^2 \leq \frac{\gamma_n \|\tilde{s}_n^{\gamma_n}\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i \big(\tilde{s}_n^{\gamma_n}(x_i) - \hat{f}^{(n)}(x_i)\big)^2}{\gamma_n \mu_n^{\gamma_n}(y)}. \tag{46}$$

The right-hand side of inequality (46) can further be bounded as follows. On the one hand, the optimality of the approximant $\tilde{s}_n^{\gamma_n}$ provides

$$\begin{aligned}
\gamma_n \|\tilde{s}_n^{\gamma_n}\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i \big(\tilde{s}_n^{\gamma_n}(x_i) - \hat{f}^{(n)}(x_i)\big)^2 &\\
&\leq \gamma_n \|\tilde{s}_n\|_\phi^2 + \frac{1}{n}\sum_{i=1}^n w_i \big(\tilde{s}_n(x_i) - \hat{f}^{(n)}(x_i)\big)^2 \\
&\leq \gamma_n \big\|f\big\|_{\mathcal{N}_\phi}^2 + \frac{1}{n}\sum_{i=1}^n w_i \big(f(x_i) - \hat{f}^{(n)}(x_i)\big)^2,
\end{aligned} \tag{47}$$

where $\tilde{s}_n$ is the optimal interpolant to the data $(x_i, f(x_i))$, $i = 1, \dots, n$, and $(y, f(y))$, whose semi-norm is bounded by $\|f\|_{\mathcal{N}_\phi}$ as $f \in \mathcal{N}_\phi(\mathcal{X})$, see Definition 1. On the other hand, we have for sufficiently large $n \geq n_0$ with $n\gamma_n \leq n_0\gamma_0$ that

$$\begin{aligned}
\mu_n^{\gamma_n}(y) &\geq \|l_n^{\gamma_n}(y, \cdot)\|_\phi^2 + \frac{1}{n_0\gamma_{n_0}}\sum_{i=1}^{n_0} w_i \big(l_n^{\gamma_n}(y, x_i)\big)^2 \\
&\geq \|l_{n_0}^{\gamma_{n_0}}(y, \cdot)\|_\phi^2 + \frac{1}{n_0\gamma_{n_0}}\sum_{i=1}^{n_0} w_i \big(l_{n_0}^{\gamma_{n_0}}(y, x_i)\big)^2 = \mu_{n_0}^{\gamma_{n_0}}(y),
\end{aligned}$$

where $l_{n_0}^{\gamma_{n_0}}(y, \cdot)$ with regularisation parameter $\gamma_{n_0} > 0$ is the optimal approximant to $(x_1, 0), \dots, (x_{n_0}, 0)$ with respective weights $w_1, \dots, w_{n_0}$, and subject to $l_{n_0}^{\gamma_{n_0}}(y, y) = 1$. By Cramer's rule, the positive function $\mu_{n_0}^{\gamma_{n_0}}$ can then be computed as $\mu_{n_0}^{\gamma_{n_0}}(y) = \det A_{n_0}^{\gamma_{n_0}} / \det A_{n_0}^{\gamma_{n_0}}(y)$, where the nonsingular matrices $A_{n_0}^{\gamma_{n_0}}$ and $A_{n_0}^{\gamma_{n_0}}(y)$ are given in (36) for $n = n_0$, respectively. Now, $\det A_{n_0}^{\gamma_{n_0}}$ is a nonzero

constant and $\det A_{n_0}^{\gamma_{n_0}}(y)$ is bounded on $\mathcal{X}$, as a continuous function. It thus follows that $\mu_{n_0}^{\gamma_{n_0}}(y)$ is bounded away from zero. Hence, there exists a constant $\alpha_1 > 0$, depending on $x_1, \ldots, x_{n_0}$ and on $\gamma_{n_0}$, such that

$$\mu_n^{\gamma_n}(y) \geq \alpha_1, \qquad \forall\, y \in \mathcal{X}\backslash\{x_1, \ldots, x_n\},\ n \geq n_0. \tag{48}$$

Consequently, by (47) and (48), we get that inequality (46) reduces to

$$\big[f(y) - s_n^{\gamma_n}(y)\big]^2 \leq \frac{\|f\|_{\mathcal{N}_\phi}^2 + \frac{1}{n\gamma_n}\sum_{i=1}^n w_i\big(f(x_i) - \hat{f}^{(n)}(x_i)\big)^2}{\alpha_1},$$

which, as $f$ is bounded on $\mathcal{X}$, results in

$$\big|s_n^{\gamma_n}(y)\big| \leq \frac{1}{\sqrt{\alpha_1}}\Big(\|f\|_{\mathcal{N}_\phi}^2 + \frac{1}{n\gamma_n}\sum_{i=1}^n w_i\big(f(x_i) - \hat{f}^{(n)}(x_i)\big)^2\Big)^{1/2} + \|f\|_\infty, \tag{49}$$

for $y \in \mathcal{X}\backslash\{x_1, \ldots, x_n\}$.

Due to the continuous extension of $\mu_n^{\gamma_n}$, inequality (48) also applies at the sample points, cf. equation (37). Accordingly, since $\phi$ is assumed to be conditional positive definite and thus continuous, the upper bound in (49) is also valid for $s_n^{\gamma_n}$ at the sample points $x_1, \ldots, x_n$. □

Under additional assumptions on the scaled weighted sum of squared errors in inequality (43) such that $\{\|s_n^{\gamma_n}\|_\infty\}$ is bounded uniformly, the following convergence result for sufficiently smooth objective functions $f$ can then be established together with Theorems 2 and 5. For the analogous deterministic case, see Gutmann (2001a), Corollary 11 (or Gutmann (2001c), Corollary 4.8).

**Corollary 2** *Let $\phi$ and $m$ be as in Theorem 5. Further, let $\nu_d$ be as in Theorem 4, $f \in C^{\nu_d}(\mathcal{X})$ with minimal function value $f^*$, and let the sequence $(1/n\gamma_n)\sum_{i=1}^n w_i(f(x_i) - \hat{f}^{(n)}(x_i))^2\}$ be convergent. Assume that, for infinitely many $n \in \mathbb{N}$, we have*

$$f_n^* < \min_{y\in\mathcal{X}}\Big[s_n^{\gamma_n}(y) - \tau\big[\Delta_n(y) + \widetilde{w}_n^{-1/2}(y)\big]^{\rho/2}\Big],$$

*where $\tau$, $\Delta_n$, $\rho$ and $\widetilde{w}_n$ are given as above, and that $n\gamma_n \to 0$ as $n \to \infty$. Then, we have $\lim_{n\to\infty} f(x^{(n)}) = f^*$ for the sequence $\{x^{(n)}\}$ constructed by Algorithm 2.*

*Remark 4* Note that Lemma 1 may also be formulated for noisy functions $\hat{f}$ in the native space, i.e. for functions with sufficiently well-behaved noise. In this case, $\|s_n^{\gamma_n}\|_\infty$ can be bounded uniformly by a number that only depends on $x_1, \ldots, x_{n_0}$, $\gamma_{n_0}$, and $\hat{f}$, such that Corollary 2 holds for $\hat{f} \in \mathcal{N}_\phi(\mathcal{X})$.

### 5.3 Proof of Convergence

To prove Theorem 5, we require some lemmas on the behaviour of the functions $\mu_n^{\gamma_n}$, $n \in \mathbb{N}$. The lemmas essentially generalise Lemmas 12 – 14 of Gutmann (2001a) (or Lemmas 4.9–4.11 in Gutmann (2001c)) in order to account for the presence of noise. Correspondingly, the first two lemmas are concerned with the limit of the sequence $\{\mu_n^{\gamma_n}(x_n)\}$.

**Lemma 2** *Let $\phi$ be a conditionally positive definite radial basis function of order $m$ from Table 1, and let $\{z_1, \dots, z_k\}$ be a $\mathcal{P}_m^d$-unisolvent set in a compact set $\mathcal{X} \subset \mathbb{R}^d$. Let $\{x_n\}$ and $\{y_n\}$ be convergent sequences in $\mathcal{X}$ that have the same limit $x^* \notin \{z_1, \dots, z_k\}$ and satisfy $x_n \neq y_n$, $n \in \mathbb{N}$. Further, let $\tilde{l}_n^{\gamma_n}(x_n, \cdot)$ with $\gamma_n > 0$ be the optimal regularised least-squares approximant to the data $(z_1, 0), \dots, (z_k, 0), (y_n, 0)$ and subject to $\tilde{l}_n^{\gamma_n}(x_n, x_n) = 1$, where the corresponding weights $w_1, \dots, w_k, w_n$ are bounded away from zero. If $n\gamma_n \to 0$ as $n \to \infty$, then*

$$\lim_{n\to\infty} \left[\|y_n - x_n\|_2 + w_n^{-1/2}\right]^\rho \tilde{\mu}_n^{\gamma_n}(x_n) = \infty, \tag{50}$$

*where $\tilde{\mu}_n^{\gamma_n}$ is the function defined by (35) for the approximant $\tilde{l}_n^{\gamma_n}(x_n, \cdot)$, and where $0 \leq \rho < 1$, for $\phi(r) = r$, and $0 \leq \rho < 2$, otherwise.*

*Proof* For $\gamma_n > 0$, consider the optimal approximant $\tilde{l}_n^{\gamma_n}(x_n, \cdot)$ to $(z_1, 0), \dots, (z_k, 0), (y_n, 0)$, with corresponding weights $w_1, \dots, w_k, w_n$, and interpolating $(x_n, 1)$. For sufficiently large $n$, neither $x_n$ nor $y_n$ is in the set $\{z_1, \dots, z_k\}$, so that Cramer's rule may be applied to compute the function $\tilde{\mu}_n^{\gamma_n}$ associated to $\tilde{l}_n^{\gamma_n}(x_n, \cdot)$ by

$$\tilde{\mu}_n^{\gamma_n}(x_n) = \frac{\det A_n^{\gamma_n}}{\det A_n^{\gamma_n}(x_n)},$$

where the nonsingular matrices $A_n^{\gamma_n}$ and $A_n^{\gamma_n}(x_n)$ are of the form (36) for the points $z_1, \dots, z_k, y_k$ and $z_1, \dots, z_k, y_k, x_n$, respectively. In particular, the latter matrix can be written as

$$A_n^{\gamma_n}(x_n) = \begin{pmatrix} \Phi + n\gamma_n W^{-1} & u_k(y_n) & u_k(x_n) & P \\ u_k(y_n)^\top & \phi(0) + n\gamma_n w_n^{-1} & \phi(\|y_n - x_n\|_2) & \pi(y_n)^\top \\ u_k(x_n)^\top & \phi(\|y_n - x_n\|_2) & \phi(0) & \pi(x_n)^\top \\ P^\top & \pi(y_n) & \pi(x_n) & 0 \end{pmatrix},$$

where $\Phi \in \mathbb{R}^{k\times k}$ and $P \in \mathbb{R}^{k\times \widetilde{m}}$ correspond to the interpolation and polynomial basis matrix of $\{z_1, \dots, z_k\}$, respectively, $W = \operatorname{diag}(w_1, \dots, w_k)$, and $u_k(y) = (\phi(\|z_1 - y\|_2), \dots, \phi(\|z_k - y\|_2))^\top$ and $\pi(y) = (p_1(y), \dots, p_{\widetilde{m}}(y))^\top$ for any $y \in \mathcal{X}$.

By the continuity of the determinant and the assumption $n\gamma_n \to 0$ as $n \to \infty$ with weights bounded away from zero, it follows that $\lim_{n\to\infty} \det A_n^{\gamma_n} = \det A^* \neq 0$, where $A^*$ denotes the nonsingular interpolation matrix given in form of the left-hand side of (8) for the points $z_1, \dots, z_k, x^*$. In order to show assertion (50), it therefore remains to consider expression

$$\left[\|y_n - x_n\|_2 + w_n^{-1/2}\right]^{-\rho} \det A_n^{\gamma_n}(x_n), \tag{51}$$

for which we show in the following that it converges to zero as $n \to \infty$. First note that the $(k+1)$-th and $(k+2)$-th rows of the matrix $A_n^{\gamma_n}(x_n)$, given by

$$\begin{aligned} &\left(u_k(y_n)^\top \;\; \phi(0) + n\gamma_n w_n^{-1} \;\; \phi(\|y_n - x_n\|_2) \;\; \pi(y_n)^\top\right), \text{ and} \\ &\left(u_k(x_n)^\top \;\; \phi(\|y_n - x_n\|_2) \qquad \phi(0) \qquad \pi(x_n)^\top\right), \end{aligned}$$

have the same limit for $n \to \infty$, as the weights are bounded away from zero and $n\gamma_n \to 0$ for $n \to \infty$. Consequently, $\det A_n^{\gamma_n}(x_n) \to 0$ as $n \to \infty$, and hence, for $\rho = 0$, assertion (50) follows immediately.

For $\rho > 0$, note that the determinant of $A_n^{\gamma_n}(x_n)$ does not change if the $(k+1)$-th row of the matrix $A_n^{\gamma_n}(x_n)$ is replaced by the difference between the $(k+1)$-th and the $(k+2)$-th row, and, subsequently, the $(k+1)$-th column is replaced by the difference between the $(k+1)$-th and the $(k+2)$-th column. Therefore, $\det A_n^{\gamma_n}(x_n)$ can equally be computed as

$$\begin{vmatrix} \Phi + n\gamma_n W^{-1} & u_k(y_n) - u_k(x_n) & u_k(x_n) & P \\ u_k(y_n)^\top - u_k(x_n)^\top & 2[\phi(0) - \phi(\|y_n - x_n\|_2)] + n\gamma_n w_n^{-1} & \phi(\|y_n - x_n\|_2) - \phi(0) & \pi(y_n)^\top - \pi(x_n)^\top \\ u_k(x_n)^\top & \phi(\|y_n - x_n\|_2) - \phi(0) & \phi(0) & \pi(x_n)^\top \\ P^\top & \pi(y_n) - \pi(x_n) & \pi(x_n) & 0 \end{vmatrix}.$$

To deduce the convergence of expression (51) to zero, we then divide the $(k+1)$-th row and the $(k+1)$-th column of the latter determinant by $[\|y_n - x_n\|_2 + w_n^{-1/2}]^{\rho/2}$, and make the following remarks on the newly formed $(k+1)$-th column.

For all choices of $\phi$, the functions $\phi(\|z_i - \cdot\|_2)$, $i = 1, \ldots, k$, are Lipschitz continuous on $\mathcal{X}$. This implies for $\rho < 2$ that

$$\lim_{n\to\infty} \frac{\phi(\|z_i - y_n\|_2) - \phi(\|z_i - x_n\|_2)}{\left[\|y_n - x_n\|_2 + w_n^{-1/2}\right]^{\rho/2}} = 0, \qquad i = 1, \ldots, k,$$

such that $u_k(y_n) - u_k(x_n) \to 0$ as $n \to \infty$. Similarly, for the same choice of $\rho$, the Lipschitz continuity of the polynomials yields

$$\lim_{n\to\infty} \frac{p_j(y_n) - p_j(x_n)}{\left[\|y_n - x_n\|_2 + w_n^{-1/2}\right]^{\rho/2}} = 0, \qquad j = 1, \ldots, \widetilde{m},$$

resulting in $\pi(y_n) - \pi(x_n) \to 0$ as $n \to \infty$. Further, we have

$$\lim_{n\to\infty} \frac{\phi(\|y_n - x_n\|_2) - \phi(0)}{\left[\|y_n - x_n\|_2 + w_n^{-1/2}\right]^{\rho}} = 0,$$

for $\rho < \nu$ in the case of surface splines and for $\rho < 2$ in the other cases. This follows directly in the case of surface splines, due to their form, and by the second order Taylor expansion in the other cases, as $\phi'(0) = 0$ and $\phi''(r)$ is bounded for small $r$.

Eventually, by assuming that $n\gamma_n \to 0$ as $n \to \infty$ and since $w_n$ is bounded away from zero, we observe for $\rho < 2$,

$$\lim_{n\to\infty} \frac{n\gamma_n w_n^{-1}}{\left[\|y_n - x_n\|_2 + w_n^{-1/2}\right]^{\rho}} = 0.$$

Altogether, we therefore have that expression (51) converges for the given choices of $\rho$ to zero as $n \to \infty$, proving that assertion (50) also holds in case $\rho > 0$. □

**Lemma 3** *Let $\phi$ and $m$ be as in Lemma 2, where $\rho$ takes a value as indicated. Let $\{x_n\}$ be a sequence in $\mathcal{X}$ with pairwise different points such that $\{x_1, \ldots, x_{n_0}\}$ is $\mathcal{P}_m^d$-unisolvent. For any $y \in \mathcal{X}\backslash\{x_1, \ldots, x_n\}$, let $l_n^{\gamma_n}(y, \cdot)$ with $\gamma_n > 0$ be the optimal regularised least-squares approximant to the data $(x_1, 0), \ldots, (x_n, 0)$ and subject to $l_n^{\gamma_n}(y, y) = 1$, where the corresponding weights $w_1, \ldots, w_n$ are bounded away from zero. If $n\gamma_n \to 0$ as $n \to \infty$, then for every convergent subsequence $\{x_{n_k}\}_{k\in\mathbb{N}}$ of $\{x_n\}$ it holds*

$$\lim_{k\to\infty} \left[\Delta_{n_k-1}(x_{n_k}) + \widetilde{w}_{n_k-1}^{-1/2}(x_{n_k})\right]^{\rho} \mu_{n_k-1}^{\gamma_{n_k-1}}(x_{n_k}) = \infty,$$

*where* $\mu_{n_k-1}^{\gamma_{n_k-1}}$, $\Delta_{n_k-1}$, *and* $\widetilde{w}_{n_k-1}$ *are the functions given by* (35), (41), *and* (42), *respectively, for* $n = n_k - 1$.

*Proof* For $n \geq 2$, let $i(x_n) = \arg\min_{1\leq i\leq n-1}\|x_n - x_i\|_2$, where we choose the largest $i(x_n)$ among the minimising indices if arg min is not unique, and let the sequence $\{y_n\}_{n\in\mathbb{N}}$ be defined as

$$y_n := \begin{cases} x_2, & n = 1, \\ x_{i(x_n)}, & n \geq 2. \end{cases}$$

Further, let $\{x_{n_k}\}$ be any subsequence of $\{x_n\}$ that converges to a point $x^* \in \mathcal{X}$. The choice of $\{y_n\}$ and convergence thus yield $\lim_{k\to\infty}\|x_{n_k} - y_{n_k}\|_2 = 0$. Also note that there always exists a $\mathcal{P}_m^d$-unisolvent set $\{\bar{x}_1, \ldots, \bar{x}_l\}$, $l \in \mathbb{N}$, in the sequence $\{x_n\}$ that does not contain the limit point $x^*$. If $x^* = x_i$ for some $i \in \{1, \ldots, n_0\}$, then we can pick $x_{n_i}$ in a neighbourhood of $x^*$ such that the initial set $\{x_1, \ldots, x_{i-1}, x_{n_i}, x_{i+1}, \ldots, x_{n_0}\}$ is $\mathcal{P}_m^d$-unisolvent.

For sufficiently large $k \in \mathbb{N}$ such that $y_{n_k} \notin \{\bar{x}_1, \ldots, \bar{x}_l\}$ and for any $y \in \mathcal{X}\backslash\{x_1, \ldots, x_{n_k-1}\}$, let $\bar{l}_k^{\gamma_k}(y, \cdot)$ with $\gamma_k > 0$ be the optimal regularised least-squares approximant to the data $(\bar{x}_1, 0), \ldots, (\bar{x}_l, 0)$, $(y_{n_k}, 0)$, with corresponding weights $\overline{w}_1, \ldots, \overline{w}_l$, $w_{n_k}$ bounded away from zero, and subject to $\bar{l}_k^{\gamma_k}(y, y) = 1$. Likewise, let $l_{n_k-1}^{\gamma_{n_k-1}}(y, \cdot)$ with $\gamma_{n_k-1} > 0$ be the optimal regularised least-squares approximant to $(x_1, 0), \ldots, (x_{n_k-1}, 0)$, with corresponding weights $w_1, \ldots, w_{n_k-1}$ bounded away from zero, and subject to $l_{n_k-1}^{\gamma_{n_k-1}}(y, y) = 1$. Observe that, for $k$ large enough, $l_{n_k-1}^{\gamma_{n_k-1}}(y, \cdot)$ approximates $(\bar{x}_i, 0)$, $i = 1, \ldots, l$, and $(y_{n_k}, 0)$, along with their given weights and subject to the same interpolation condition. Hence, for sufficiently large $k$, the functions $\bar{\mu}_k^{\gamma_k}$ and $\mu_{n_k-1}^{\gamma_{n_k-1}}$ associated to $\bar{l}_k^{\gamma_k}(y, \cdot)$ and $l_{n_k-1}^{\gamma_{n_k-1}}(y, \cdot)$ via (35), respectively, and the optimality of $\bar{l}_k^{\gamma_k}(y, \cdot)$ imply

$$\begin{aligned} \bar{\mu}_k^{\gamma_k}(y) &= \big\|\bar{l}_k^{\gamma_k}(y, \cdot)\big\|_\phi^2 + \frac{1}{(l+1)\gamma_k}\left[\sum_{i=1}^{l} \overline{w}_i\big(\bar{l}_k^{\gamma_k}(y, \bar{x}_i)\big)^2 + w_{n_k}\big(\bar{l}_k^{\gamma_k}(y, y_{n_k})\big)^2\right] \\ &\leq \big\|l_{n_k-1}^{\gamma_{n_k-1}}(y, \cdot)\big\|_\phi^2 + \frac{1}{(l+1)\gamma_k}\sum_{i=1}^{n_k-1} w_i\big(l_{n_k-1}^{\gamma_{n_k-1}}(y, x_i)\big)^2 \\ &\leq \mu_{n_k-1}^{\gamma_{n_k-1}}(y), \end{aligned} \tag{52}$$

where the last inequality follows from the assumption that $n\gamma_n \to 0$ as $n \to \infty$.

Eventually, by definition of the sequence $\{y_n\}$ and applying Lemma 2 with the set of points $\{z_1, \ldots, z_k\}$ being $\{\bar{x}_1, \ldots, \bar{x}_l\}$, the weights $w_1 \ldots, w_k$ being replaced by $\overline{w}_1, \ldots, \overline{w}_l$, and setting $n = n_k$, we obtain

$$\begin{aligned} &\lim_{k\to\infty}\big[\Delta_{n_k-1}(x_{n_k}) + \widetilde{w}_{n_k-1}^{-1/2}(x_{n_k})\big]^\rho\, \bar{\mu}_k^{\gamma_k}(x_{n_k}) \\ &\qquad = \lim_{k\to\infty}\big[\|x_{n_k} - y_{n_k}\|_2 + w_{n_k}^{-1/2}\big]^\rho\, \bar{\mu}_k^{\gamma_k}(x_{n_k}) = \infty, \end{aligned}$$

for the given choice of $\rho$. Consequently, by setting $y = x_{n_k}$ in (52), it follows that $[\Delta_{n_k-1}(x_{n_k}) + \widetilde{w}_{n_k-1}^{-1/2}(x_{n_k})]^\rho\, \mu_{n_k-1}^{\gamma_{n_k-1}}(x_{n_k})$ tends to infinity for $k \to \infty$, as claimed. □

Akin to Gutmann (2001a), Lemma 14 (or Gutmann (2001c), Lemma 4.11), the next lemma states that the sequence $\{\mu_n^{\gamma_n}(y)\}$ is uniformly bounded if $y$ is bounded away from the points in the sequence $\{x_n\}$. Note that this result only holds for surface splines as Theorem 4 is required.

**Lemma 4** *Let $\phi$ be a conditionally positive definite surface spline of order $m$ from Table 1, and let $\{x_n\}$ be a sequence in $\mathbb{R}^d$ with pairwise different points such that $\{x_1, \ldots, x_{n_0}\}$ is $\mathcal{P}_m^d$-unisolvent. Further, let $y_0 \in \mathbb{R}^d$ satisfy $\|y_0 - x_n\|_2 \geq \delta$, $n \in \mathbb{N}$, for some $\delta > 0$. Then, there exists $C > 0$, depending only on $y_0$ and $\delta$, such that*

$$\mu_n^{\gamma_n}(y_0) \leq C, \qquad \forall\, n \geq n_0,$$

*where $\mu_n^{\gamma_n}$ with $\gamma_n > 0$ is the function given by* (35).

*Proof* Let $B_\delta(y_0) = \{x \in \mathbb{R}^d : \|x - y_0\|_2 < \delta\}$. There exists a compactly supported function $\varphi \in C^\infty(\mathbb{R}^d)$ that takes the value 1 at $y_0$ and 0 on $\mathbb{R}^d \backslash B_\delta(y_0)$. It follows from Theorem 4 that $\varphi \in \mathcal{N}_\phi(\mathbb{R}^d)$.

For any $n \geq n_0$, let $l_n(y_0, \cdot)$ be the optimal interpolant to the data $(x_1, 0), \ldots, (x_n, 0)$ and $(y_0, 1)$, such that $l_n(y_0, x_i) = \varphi(x_i) = 0$, $i = 1, \ldots, n$, and $l_n(y_0, y_0) = \varphi(y_0) = 1$. Similarly, for any $n \geq n_0$, let $l_n^{\gamma_n}(y_0, \cdot)$ with $\gamma_n > 0$ denote the optimal regularised least-squares approximant to $(x_1, 0), \ldots, (x_n, 0)$, with corresponding weights $w_1, \ldots, w_n$, and subject to $l_n^{\gamma_n}(y_0, y_0) = 1$. By definition of $\mu_n^{\gamma_n}$ and the optimality of $l_n^{\gamma_n}(y_0, \cdot)$, we then have

$$\mu_n^{\gamma_n}(y_0) \leq \|l_n(y_0, \cdot)\|_\phi^2 + \frac{1}{n\gamma_n} \sum_{i=1}^{n} w_i \big(l_n(y_0, x_i)\big)^2 = \|l_n(y_0, \cdot)\|_\phi^2,$$

which is bounded by $C := \|\varphi\|_{\mathcal{N}_\phi}^2$, see Definition 1. □

By using the lemmas above, we can now provide the main proof of Theorem 5, stating that the sequence generated by Algorithm 2 is dense in $\mathcal{X}$. Because of the established similarity of the algorithm to Gutmann's RBF method, the proof follows the main lines of the proof of Theorem 7 in Gutmann (2001a) (or Gutmann (2001c), Theorem 4.5).

*Proof of Theorem 5* Assume that there is $y_0 \in \mathcal{X}$ and $\delta > 0$, such that $B_\delta(y_0) = \{x \in \mathbb{R}^d : \|x - y_0\|_2 < \delta\}$ does not contain any $x_n$, $n \in \mathbb{N}$. According to the iteration step of Algorithm 2, it then holds

$$g_n^{\gamma_n}(x_{n+1}) \leq g_n^{\gamma_n}(y_0), \qquad n \geq n_0,$$

where $\gamma_n > 0$. Moreover, since $f_n^*$ is assumed to satisfy condition (40) for infinitely many $n \in \mathbb{N}$, there exists a subsequence $\{x_{n_k}\}_{k \in \mathbb{N}}$ such that

$$s_{n_k-1}^{\gamma_{n_k-1}}(x_{n_k}) - f_{n_k-1}^* > \tau \big\| s_{n_k-1}^{\gamma_{n_k-1}} \big\|_\infty \big[\Delta_{n_k-1}(x_{n_k}) + \widetilde{w}_{n_k-1}^{-1/2}(x_{n_k})\big]^{\rho/2},$$

for the specified quantities $\tau$ and $\rho$, and the functions $\Delta_{n_k-1}$ and $\widetilde{w}_{n_k-1}$ as given by (41) and (42), respectively, for $n = n_k - 1$. Now, the sequence $\{x_{n_k}\}$ has a convergent subsequence which, without loss of generality, shall be denoted again by $\{x_{n_k}\}$. Since each $x_{n_k}$, $k \in \mathbb{N}$, minimises $g_{n_k-1}$ on $\mathcal{X} \backslash \{x_1, \ldots, x_{n_k-1}\}$,

the same reasoning as in the proof of Theorem 7 in Gutmann (2001a), inequality (A.11) to (A.12), then leads to the inequality

$$\mu_{n_k-1}^{\gamma_{n_k-1}}(x_{n_k})\big[\Delta_{n_k-1}(x_{n_k}) + \widetilde{w}_{n_k-1}^{-1/2}(x_{n_k})\big]^{\rho} \\ \leq \mu_{n_k-1}^{\gamma_{n_k-1}}(y_0)\left[\big[\Delta_{n_k-1}(x_{n_k}) + \widetilde{w}_{n_k-1}^{-1/2}(x_{n_k})\big]^{\rho/2} + \frac{2}{\tau}\right]^2, \tag{53}$$

which renders a contradiction by virtue of Lemmas 2 – 4. In particular, on the one hand, Lemma 3 reveals that the left-hand side of (53) converges to infinity for $k \to \infty$. On the other hand, Lemma 4 shows that $\mu_n^{\gamma_n}(y_0)$ is bounded above by some constant independent of $n$, which together with the uniform boundedness of $\Delta_{n_k-1}(x_{n_k})$ on $\mathcal{X}$ and the weights being bounded away from zero implies that the right-hand side of inequality (53) is bounded above by a constant independent of $k$. Hence, due to this contradiction, we can deduce that $B_\delta(y_0)$ must contain a point of the sequence $\{x_n\}$, so that, eventually, $\{x_n\}$ is dense in the compact set $\mathcal{X}$. □

## 6 Conclusions

In this paper, we have addressed the global optimisation of an expensive and noisy objective function where observed function values are assumed to lie within error bounds. Based on Gutmann's original RBF method for minimising a deterministic objective function, relying on radial basis function interpolation, we have first discussed common approaches of radial basis function approximations for integration into a response surface method. Arguing in favour of regularised least-squares approximants, we then have presented a noisy RBF method that constructs the smoothest possible response surfaces that stay within the given error bounds at the evaluated points, and determines new evaluation points by minimising a regularised least-squares criterion in terms of a target value. Further on, we have established convergence of the noisy RBF method to the global minimum of any continuous function, under some additional assumptions on the error bounds, and provided relevant convergence results. Future work will include the assessment of the proposed method on various academic and real-world test functions.

## 7 Compliance with Ethical Standards

This work was supported by the Engineering and Physical Sciences Research Council [grant no. EP/M50662X/1] of the UK.

The authors declare that there are no conflicts of interests or competing interests.

## References

Akhtar T, Shoemaker CA (2016) Multi objective optimization of computationally expensive multi-modal functions with rbf surrogates and multi-rule selection. Journal of Global Optimization 64:17–32

Björkman M, Holmström K (2000) Global Optimization of Costly Nonconvex Functions Using Radial Basis Functions. Optimization and Engineering 1(4):373–397

Buhmann MD (2003) Radial Basis Functions: Theory and Implementations. Cambridge University Press, Cambridge

Cassioli A, Schoen F (2013) Global optimization of expensive black box problems with a known lower bound. Journal of Global Optimization 57(1):177–190

Costa A, Nannicini G (2018) RBFOpt: an open-source library for black-box optimization with costly function evaluations. Mathematical Programming Computation 10(4):597–629

Ehrgott M (2005) Multicriteria Optimization. Springer, Berlin Heidelberg

Feliot P, Bect J, Vazquez E (2017) A bayesian approach to constrained single-and multi-objective optimization. Journal of Global Optimization 67(1-2):97–133

Forrester AIJ, Keane AJ (2009) Recent advances in surrogate-based optimization. Progress in Aerospace Sciences 45(1–3):50–79

Fowkes J (2011) Bayesian Numerical Analysis: Global Optimization and Other Applications. PhD thesis, University of Oxford

Gramacy RB, Lee HKH (2011) Optimization Under Unknown Constraints. In: Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M (eds) Bayesian Statistics 9, Oxford University Press, Oxford

Gutmann HM (2001a) A Radial Basis Function Method for Global Optimization. Journal of Global Optimization 19(3):201–227

Gutmann HM (2001b) On the semi-norm of radial basis function interpolants. Journal of Approximation Theory 111(2):315–328, DOI https://doi.org/10.1006/jath.2001.3579, URL https://www.sciencedirect.com/science/article/pii/S0021904501935798

Gutmann HM (2001c) Radial basis function methods for global optimization. PhD thesis, University of Cambridge

Habib A, Singh HK, Ray T (2016) A study on the effectiveness of constraint handling schemes within efficient global optimization framework. In: 2016 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, pp 1–8

Holmström K (2008) An adaptive radial basis algorithm (ARBF) for expensive black-box global optimization. Journal of Global Optimization 41(3):447–464

Huang D, Allen TT, Notz WI, Zeng N (2006) Global Optimization of Stochastic Black-Box Systems via Sequential Kriging Meta-Models. Journal of Global Optimization 34(3):441–466

Iske A (2004) Multiresolution Methods in Scattered Data Modelling. Springer, Berlin Heidelberg

Jakobsson S, Patriksson M, Rudholm J, Wojciechowski A (2010) A method for simulation based optimization using radial basis function. Optimization and Engineering 11(4):501–532

Ji Y, Kim S, Lu WX (2013) A new framework for combining global and local methods in black box optimization. Optimization Online paper 3977

Jones DR (1996) Global Optimization with Response Surfaces. Fifth SIAM Conference on Optimization, Victoria, Canada

Jones DR (2001) A Taxonomy of Global Optimization Methods Based on Response Surfaces. Journal of Global Optimization 21(4):345–383

Jones DR, Schonlau M, Welch WJ (1998) Efficient Global Optimization of Expensive Black-Box Functions. Journal of Global Optimization 13(4):455–492

Kelley CT (1999) Iterative Methods for Optimization. SIAM, Philadelphia, PA
Kushner HJ (1962) A versatile stochastic model of a function of unknown and time varying form. Journal of Mathematical Analysis and Applications 5(1):150–167
Kushner HJ (1964) A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. Journal of Basic Engineering 86(1):97–106
Matheron G (1963) Principles of geostatistics. Economic Geology 58(8):1246–1266
Mockus J, Tiesis V, Žilinskas A (1978) The Application of Bayesian Methods for Seeking the Extremum. In: Dixon LCW, GP S (eds) Towards Global Optimization 2, North-Holland, Amsterdam, pp 117–128
Regis RG (2014) Constrained optimization by radial basis function interpolation for high-dimensional expensive black-box problems with infeasible initial points. Engineering Optimization 46(2):218–243
Regis RG (2016) Multi-objective constrained black-box optimization using radial basis function surrogates. Journal of computational science 16:140–155
Regis RG, Shoemaker CA (2005) Constrained Global Optimization of Expensive Black Box Functions Using Radial Basis Functions. Journal of Global Optimization 31(1):153–171
Regis RG, Shoemaker CA (2007a) A Stochastic Radial Basis Function Method for the Global Optimization of Expensive Functions. INFORMS Journal on Computing 19(4):497–509
Regis RG, Shoemaker CA (2007b) Improved strategies for radial basis function methods for global optimization. Journal of Global Optimization 37(1):113–135
Regis RG, Shoemaker CA (2007c) Parallel radial basis function methods for the global optimization of expensive functions. European Journal of Operational Research 182(2):514–535
Regis RG, Shoemaker CA (2013a) A quasi-multistart framework for global optimization of expensive functions using response surface models. Journal of Global Optimization 56:1719–1753
Regis RG, Shoemaker CA (2013b) A quasi-multistart framework for global optimization of expensive functions using response surface models. Journal of Global Optimization 56(4):1719–1753
Sacks J, Welch WJ, Mitchell TJ, Wynn HP (1989) Design and Analysis of Computer Experiments. Statistical Science 4(4):409–423
Schaback R (1993) Comparison of Radial Basis Function Interpolants. In: Jetter K, Utreras FI (eds) Multivariate Approximation: From CAGD to Wavelets, World Scientific, Singapore, pp 293–305
Schaback R (1999) Native spaces for radial basis functions I. In: Müller MW, Buhmann MD, Mache DH, Felten M (eds) New Developments in Approximation Theory, Birkhäuser, Basel, pp 255–282
Schaback R, Wendland H (2006) Kernel techniques: From machine learning to meshless methods. Acta Numerica 15:543–639
Schölkopf B, Smola AJ (2002) Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA
Schonlau M (1997) Computer Experiments and Global Optimization. PhD thesis, University of Waterloo
Shen Y, Shoemaker CA (2020) Global optimization for noisy expensive black-box multi-modal functions via radial basis function surrogate. In: 2020 Winter Simulation Conference (WSC), IEEE, pp 3020–3031

Sóbester A, Leary SJ, Keane AJ (2005) On the Design of Optimization Strategies Based on Global Response Surface Approximation Models. Journal of Global Optimization 33(1):31–59

Törn A, Žilinskas A (1989) Global Optimization. Springer, Berlin Heidelberg

Villemonteix J, Vazquez E, Walter E (2009) An informational approach to the global optimization of expensive-to-evaluate functions. Journal of Global Optimization 44(4):509–534

Vu KK, D'Ambrosio C, Hamadi Y, Liberti L (2017) Surrogate-based methods for black-box optimization. International Transactions in Operational Research 24(3):393–424

Žilinskas A (2010) On similarities between two models of global optimization: statistical models and radial basis functions. Journal of Global Optimization 48(1):173–182

Wang J, Clark SC, Liu E, Frazier PI (2020) Parallel bayesian global optimization of expensive functions. Operations Research 68(6):1850–1865

Wang W, Akhtar T, Shoemaker CA (2022) Integrating $\varepsilon$-dominance and rbf surrogate optimization for solving computationally expensive many-objective optimization problems. Journal of Global Optimization 82(4):965–992

Wendland H (2005a) Computational aspects of radial basis function approximation. In: Jetter K, Buhmann MD, Haussmann W, Schaback R, Stöckler J (eds) Topics in Multivariate Approximation and Interpolation, vol 12, Elsevier, Amsterdam, pp 231–256

Wendland H (2005b) Scattered Data Approximation. Cambridge University Press, Cambridge

Wendland H, Rieger C (2005) Approximate Interpolation with Applications to Selecting Smoothing Parameters. Numerische Mathematik 101(4):729–748