The use of Kriging within the SCF procedure

Christian L. Ritterhoff

August-January 2019/20

1 Introduction

In today's world of chemistry using quantum mechanical calculations for analysing reactions and predicting reaction pathways has become a standard procedure. Calculations are often used to support the experimental data found as well as to rationalize proposed reaction pathways. Furthermore notions exist, that calculations could to some extend replace traditional chemical experimentation in the future [1]. Therefore it is required for calculations to be as quick and efficient as possible to be suitable for everyday use.

As it is a dominating part of the Schrödinger equation, and equations derived from it, the wave function plays a substantial role in this process. Most methods in quantum chemistry start with an optimization of the given systems wave function with respect to it's energy, making this one of the most basic steps in quantum chemical calculations. Thus, improvements in this field are greatly beneficial in the above mentioned endeavor.

Wave function optimization is often done via a self-consistent field (SCF) process, employing an iterative procedure to optimize an eigenfunction until self-consistency is reached. Since each iteration in the iterative procedure takes time, minimizing the number of iterations needed for the algorithm to converge is a key challenge [2]. Several widely accepted strategies for acceleration of this process exist [3], each with different advantages and disadvantages.

Exploring a different approach, this report introduces the so-called Kriging procedure [4] into wave function optimization. This process is a strategy using a surrogate model. This has already been successfully explored in geometry optimization [5, 6]. Here, each specific geometry of the molecule can be attributed a specific energy. Portraying this within a coordinate system, each point relates to a specific geometry and thus from these points, an energy surface can be built. One then tries to find the minimum on this energy surface in order to optimize the geometry in respect to it's energy. However, evaluating the real energy surface is expensive since it is accompanied by

extensive quantum mechanical calculations. Thus, a model of the energy surface is created, which then is a known function and therefore easily accessible. By first finding the minimum on this surrogate model, the number of steps it takes to find the minimum on the real surface is reduced. As major features exhibit similarities between the two processes, the same approach is suggested for wave function optimization [7].

After shortly touching the theory of two widely accepted methods, Kriging and its use within the SCF procedure is further explained. A short demonstration of the method is given with the example of H_2. Afterwards, optimizations of parameters and benchmarking results for a small test suite of molecules are presented, as well as advantages and disadvantages to conventional methods discussed.

2 Theory

Prior to an explanation of the methods touched in this section, a significant difference between them has to be addressed: the solution vector. While the identity of this vector is the main goal for all methods, it is differently optimized. This is shortly expanded upon below.

All of the methods discussed in this report are wave function based and as such, the Schrödinger equation

$$H\Psi = E\Psi \tag{1}$$

plays a central role. Using the variational principle, one tries to optimize the wave function in such a way, that the energy resulting from equation 1 is minimized.

Several strategies were devised in order to solve this problem and one group of these are Hartree-Fock type procedure. Here, one of the main assumptions is, that the correct wave function can be described by a single Slater determinant

$$|\Psi\rangle = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(\vec{x}_1) & \psi_2(\vec{x}_1) & \psi_3(\vec{x}_1) & \cdots & \psi_N(\vec{x}_1) \\ \psi_1(\vec{x}_2) & \psi_2(\vec{x}_2) & \psi_3(\vec{x}_2) & \cdots & \psi_N(\vec{x}_2) \\ \psi_1(\vec{x}_3) & \psi_2(\vec{x}_3) & \psi_3(\vec{x}_3) & \cdots & \psi_N(\vec{x}_3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \psi_1(\vec{x}_N) & \psi_2(\vec{x}_N) & \psi_3(\vec{x}_N) & \cdots & \psi_N(\vec{x}_N) \end{vmatrix} \tag{2}$$

The molecular orbitals ψ_i can be written as

$$\psi_i(\vec{x}_1) = \sum_\mu c_{\mu i}\phi_\mu(\vec{x}_1) \tag{3}$$

with ϕ_μ being the basis functions. These are linearly combined with the help of the molecular orbital coefficients $c_{\mu i}$. As a result, finding the correct coefficients translates into finding the correct molecular orbitals and wave function and thus the coefficients are the solution vector that is being looked for. This is e.g. the case in the DIIS procedure [7] presented in section 2.1. Obtaining a solution vector, the Fock operator and thus the respective energy can be calculated.

Finding this solution vector is always the goal of the procedure. However, there also exists a purely variational approach to optimize this vector, which doesn't require equation 3 to be used in each iteration. Instead, one works directly with the initial molecular orbitals formed with the initial molecular orbital coefficients $c^0_{\mu i}$ [8]. These are orthonormal and separated in occupied and unoccupied orbitals. Rotating an occupied and an unoccupied orbital into one another by doing a unitary matrix transformation, results in new, but still orthonormal orbitals. This can be utilized in the following way:

Based on the number of molecular orbitals, a multidimensional subspace is formed, in which each dimension is representing the correlation between an occupied and an unoccupied orbital. An arbitrary vector,

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \tag{4}$$

portraying one point in this space, can then be used as the above mentioned unitary matrix,

$$\begin{bmatrix} X & \tilde{a} & \tilde{b} \\ -\tilde{a} & X & \tilde{c} \\ -\tilde{b} & -\tilde{c} & X \end{bmatrix} \tag{5}$$

which is used to transform the initial molecular orbitals. The change from matrix to vector and vice versa is possible, because the parameters a,b and c determine the values of $\tilde{a}$,$\tilde{b}$ and $\tilde{c}$ as well as the diagonal X elements.

The orbitals, resulting from this transformation, can then be used to calculate the molecular orbital coefficients $c_{\mu i}$ which would be required to linearly build these orbitals from the existing basis functions. As in DIIS, these orbitals subsequently can be used to calculate the system's energy. As such, each point in the multidimensional subspace can be attributed a unique energy and thus it can be regarded as an energy surface. Hence, finding the energy minimum on this surface translates into finding the correct molecular orbital coefficients $c_{\mu i}$, as both correspond to the wave function with the lowest energy eigenvalue in equation 1. Thus, the vector optimized within methods using this technique, is the vector corresponding to the minimum on the above

mentioned energy surface. Both RS-RFO [9] (section 2.2) and Kriging (section 2.3) are methods, which utilize this alternative optimization route.

In the following subsections, two widely used methods for accelerated optimization are explored. An introduction into Kriging and its use is given thereafter.

2.1 The direct inversion in the iterative subspace (DIIS) procedure

An intuitive way to reduce the number of steps and thus iterations needed, is to have a better prediction of the solution vector. Based on this thought, Pulay *et al.* [7] proposed the DIIS procedure, which stores information from past iterations in order to better predict the next trial vector.

$$\mathbf{c}_{i+1} = \sum_{i=1}^{m} p_i \mathbf{c}_i \tag{6}$$

Here, $\mathbf{c}_i$ are the previous solution vectors which are linearly combined in equation 6. Using the linear combination coefficients p_i, one thus obtains the next solution vector $\mathbf{c}_{i+1}$ from a combination of the already previously existing m solution vectors $\mathbf{c}_i$. For calculating the combination coefficients p_i the following equations need to be fulfilled.

$$\sum_{i=1}^{m} p_i = 1 \tag{7}$$

$$\Delta\mathbf{c} = \sum_{i=1}^{m} p_i \left(\mathbf{c}_{i+1} - \mathbf{c}_i\right) = \mathbf{0} \tag{8}$$

Equation 8 requires the linear combination of the residuum vector of each iteration to be zero while equation 7 forces the coefficients to sum to 1 i.e. the molecular orbitals are normalized. The thus following system of linear equations can be solved using a Lagrangian type procedure yielding the coefficients p_i which then can be used in equation 6 to approximate the next trial vector [7].

This method proved hugely successful and thus has been thoroughly investigated and expanded [10, 11, 12, 13]. Over the years it has been applied to many different types of calculations (e.g. coupled-cluster models [13]), its equations have been solved in a number of different ways [11] and further improvement spawned new convergence techniques based on the DIIS approach (e.g. EDIIS [10]). Major differences are e.g. the definition of the residium vector in equation 8. Some advantages of the method are fast convergence and the fact that the approximate Hessian does not need to be stored throughout the calculation, saving huge amounts of memory in the process [7].

However, as Le Bris and coworkers [14] mention, it can still take a high number of iterations before convergence is reached which subsequently consumes a lot of memory. They also stress the need for understanding, that this is a purely ad hoc type of procedure driven by the desire for faster convergence which is still not rigorously mathematically explainable. Thus, it is not trivial to improve the convergence behaviour of the existing methods.

2.2 The restricted-step rational function optimization (RS-RFO) procedure

As mentioned in the beginning of section 2, the RS-RFO procedure only indirectly optimizes the molecular orbital coefficients and consequently also uses a different way to find them. Instead, the vector within the multidimensional subspace is optimized. As explained, each point, and thus vector, in this subspace can be attributed a unique energy. This creates an energy surface and the goal is to find the minimum of this surface, as it will correspond to the wave function with the lowest energy in respect to equation 1. This allows one to use the extremely successful restricted step optimizations. Since changing the vector in the subspace translates into movement on the energy surface, the trial vectors will be referred to as displacements in future sections. A short explanation of the techniques used within RS-RFO is presented below.

2.2.1 Restricted optimization

A restricted optimization in general describes the search for a stationary point on a given function while also fulfilling the constrains imposed by the restriction. In the presented case, one is looking for the minimum of the system's energy surface and thus move forward by calculating displacements. By the nature of the approximation (see section 2.2.3), the model gets less accurate the further one moves from the starting point. Therefore the optimization procedure is constrained with a maximum step length, imposing a restriction. This ensures the needed degree of accuracy.

2.2.2 Surrogate models

Surrogate model procedures offer an easy and inexpensive way of exploring and predicting properties of a given function without having to actually calculate them. The process generally works as shown in figure 1.

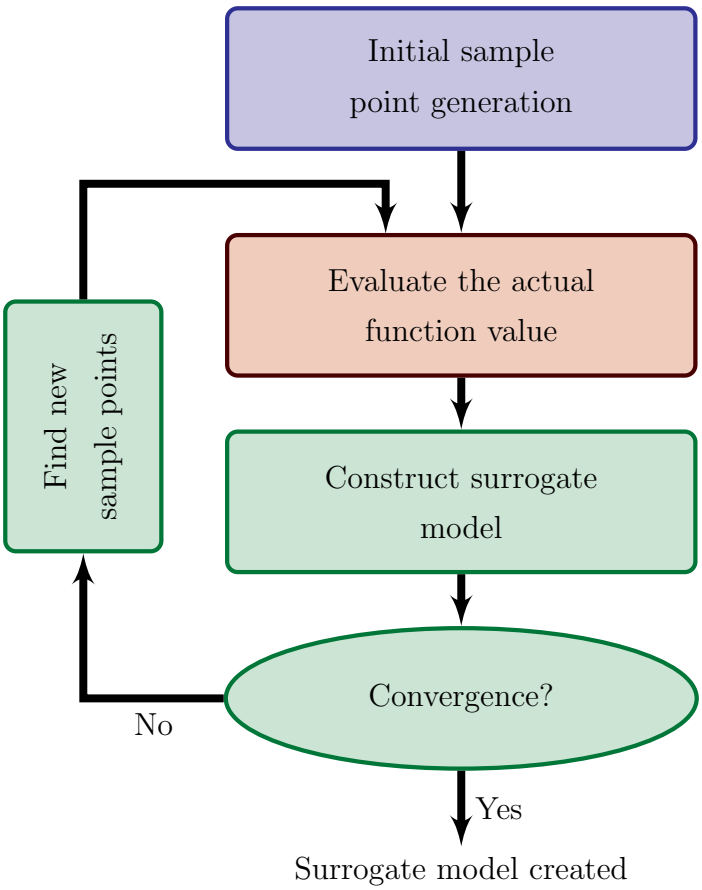


Figure 1: Sketch of the construction of a surrogate model (adapted from [15])

After initial sample points are chosen and the function value of the respective points is obtained, a surrogate is created. Convergence of this is tested by using any kind of quality test such as R^2 or Q_3 criteria [15]. Depending on the type of function/problem one faces, different strategies can be utilized to find additional sample points, should the convergence test fail. This procedure is repeated until enough data is collected for the surrogate model to fulfill the above mentioned criteria and thus convergence is reached [15].

After successful generation of a fitting model, the resulting surrogate model is a representation of the actual function. Thus it can then be used to predict the properties of this function without having to actually calculate it. Even though the creation of the model itself is time consuming, having an inexpensive approximate of the real function can easily make up for this [15]. This is especially true, when thinking about quantum mechanical calculations. Being able to calculate the energy without having to solve the one- and two-electron integrals greatly speeds up the analysation process.

2.2.3 The rational function optimisation (RFO)

Being a widely used technique to find stationary points on surfaces in general, this method can also be used for energy surfaces. It was firstly proposed by Banerjee *et al.* [16]. The search for the global minimum is begun by deviating from a starting point x_0 whilst taking small displacement steps. The neighbouring area is described by the Taylor expansion

$$E(\mathbf{x}) = E_0 + \mathbf{g}^+\mathbf{x} + \frac{1}{2}\mathbf{x}^+\mathbf{H}\mathbf{x} + ... \tag{9}$$

where $\mathbf{g}$ denotes the gradient vector and $\mathbf{H}$ represents the Hessian matrix. Second order truncation [16],

$$\epsilon = E(\mathbf{x}) - E_0 = \mathbf{g}^+\mathbf{x} + \frac{1}{2}\mathbf{x}^+\mathbf{H}\mathbf{x} \tag{10}$$

and thus approximation of this expansion is used by the Newton-Raphson procedure.

Equation 10 allows one to calculate an approximated energy for a proposed trial vector $\mathbf{x} = (x_1, ..., x_n)$ which represents a displacement from x_0. The direction of the step is chosen to be opposite to the gradient and thus towards the minimum [16]. Equation 10 can further be modified to a rational function

$$\epsilon = E(\mathbf{x}) - E_0 = \frac{\mathbf{g}^+\mathbf{x} + \frac{1}{2}\mathbf{x}^+\mathbf{H}\mathbf{x}}{1 + \mathbf{x}^+\mathbf{S}\mathbf{x}} = \frac{\frac{1}{2}(\mathbf{x}^+1)\begin{pmatrix}\mathbf{H} & \mathbf{g}\\ \mathbf{g}^+ & 0\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ 1\end{pmatrix}}{(\mathbf{x}^+1)\begin{pmatrix}\mathbf{S} & \mathbf{0}^+\\ \mathbf{0} & 1\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ 1\end{pmatrix}} \tag{11}$$

where the symmetric matrix S is most often taken as the unit matrix $\mathbf{1}$ [9]. The matrix denoted in the numerator is the so-called augmented Hessian consisting of the gradient vector and the Hessian matrix. Since this is the second order truncation of the Taylor series this represents a surrogate model with locally limited accuracy. Thus, it can only give an accurate description of the immediate area surrounding x_0, the so-called "trust region" [9]. The rational function optimization is thus controlled by a restricted-step procedure which makes sure the steps taken remain within this "trust region". Convergence is reached once a selected convergence criteria is fulfilled e.g. a certain step length or energy difference.

However, this step restriction can also present a drawback for quick convergence - the minimum is far away from the starting point of the calculation. In this case, the procedure will need a lot of iterations just to get near the quadratic region since it will not allow steps bigger than the predetermined value. This can be counteracted by dynamically adjusting the trust radius. However, no

matter how large or small the chosen step is, it always remains a guessed value. There is no way of knowing exactly how large it should be.

2.3 Kriging and its application in the SCF procedure

Kriging is a mathematical procedure originally devised by the South African mining engineer Danie G. Krige for geostatistics and was further developed by Georges Matheron (see [17] for further references). It also makes use of a surrogate model, predicting the actual function values.

2.3.1 The Kriging procedure

In the Kriging procedure, a surrogate model is constructed, predicting the energy surface of the real function and thus offering an inexpensive way to finding its minimum.

In its most general and basic approach, Kriging assumes that the estimated value $y(\mathbf{x})$ is computed as

$$y(\mathbf{x}) = \mu + \delta(\mathbf{x}) \tag{12}$$

where μ is the so-called trend function which represents the average function value, whereas $\delta(\mathbf{x})$ is the stationary covariance process [17, 4]. Therefore the correlation between the acquired data points and $\mathbf{x}$ is given by $\delta(\mathbf{x})$. Thus, $\delta(\mathbf{x})$ describes the local deviation from the constant value of the trend function based on the local influence of existing sample points.

Since calculating the actual function value is expensive, one ideally wants to use as few sample points as possible. Despite that, the obtained surrogate model should still have a high degree of accuracy. As a consequence, obtaining as much data as possible from each calculation is crucial for the quality of the model. As a result, gradient-enhanced Kriging was developed which also makes use of the gradient calculated at each sample point. This can normally be done using adjoint methods, which enriches the model with a significant amount of information at a cost much lower than a new sample point evaluation [18].

As it contains the data provided by the sample points, part of $\delta(\mathbf{x})$ is represented by a function describing the correlation between them. Numerous of these correlation functions have been suggested and discussed in literature in order to properly describe the coupling of data points (a selection of functions can be found in source [4]). In this report, a function from the Matérn [19] class of correlation functions was used, namely

$$\psi(d) = (1 + \sqrt{5}a + \frac{5a^2}{3})e^{-\sqrt{5}a} \tag{13}$$

where $d = |\mathbf{x}_{sample} - \mathbf{x}_{prediction}|$ and a is the function

$$a = \sqrt{\sum_i \frac{(x_{i,sample} - x_{i,prediction})^2}{\theta_i^2}} \tag{14}$$

depending on the hyperparameters θ_i [4], which will be explored further in future sections.

Using equations 12, 13 and 14 one can build a surrogate model in a similar fashion as described in figure 1. The resulting model is an exact interpolate of the actual function, that is to say the function values at the sample points are exactly reproduced [17]. In addition to the estimate of the real functions values, the surrogate model also provides the associated variance of these values. This is a huge advantage which will be used in a restricted variance procedure.

2.3.2 Implementation into the SCF procedure

As explained in the beginning of section 2, optimization of the wave function can be translated into finding the displacement on the energy surface, which corresponds to the lowest energy. A schematic overview of Kriging and comparison to conventional RS-RFO can be seen in figure 2.

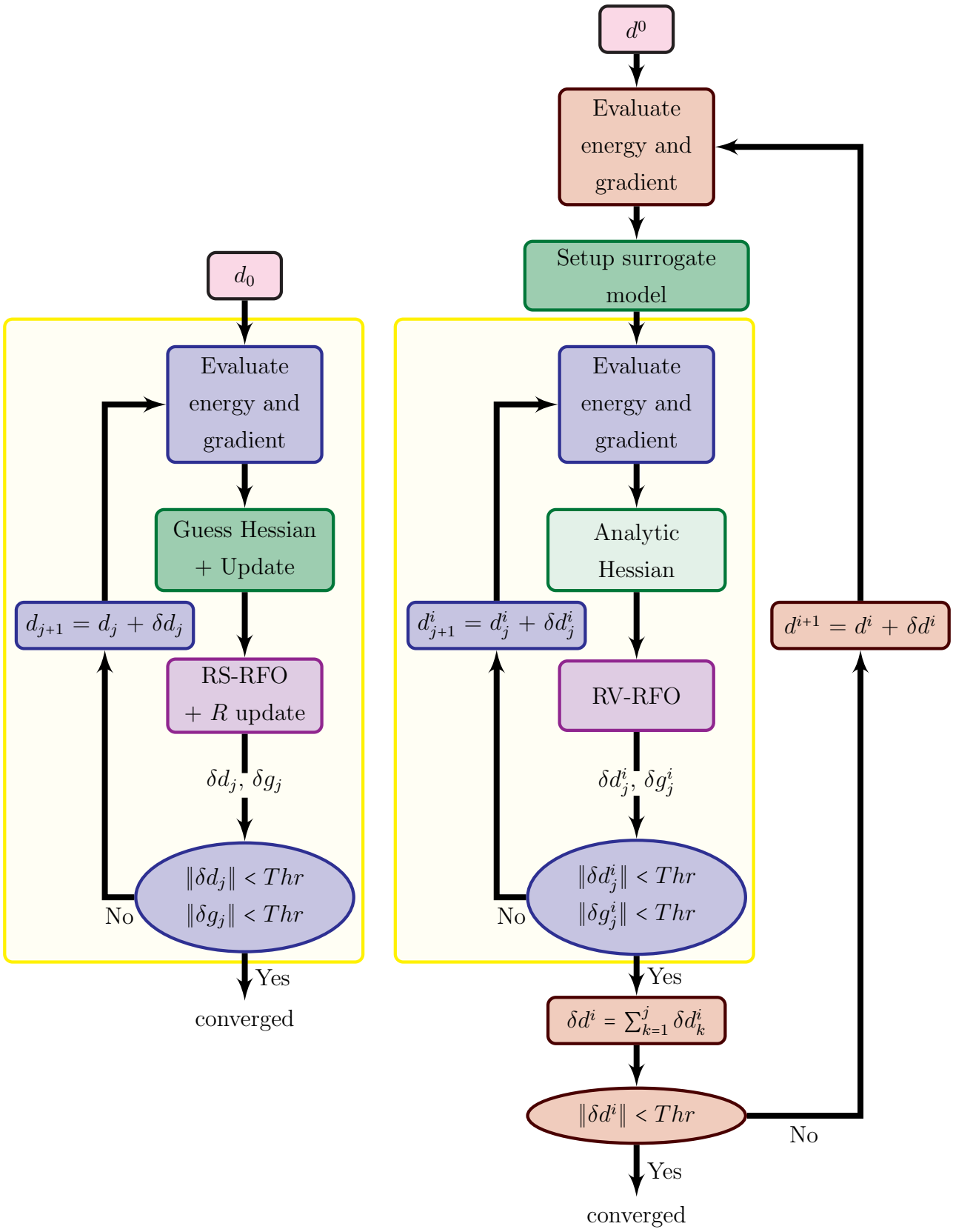


Figure 2: Schematic comparison of RS-RFO (left) and Kriging (right).

Starting with the known RS-RFO approach (left), the initial coordinates are obtained and their energy and gradient calculated. Using the guessed Hessian (green) and the RS-RFO routine (purple), a displacement is chosen. Convergence criteria then check, if a stationary point is reached. If not, the data point is evaluated and the procedure starts with the next iteration, repeating the previous steps.

Disregarding some minor differences at first, this is also what happens in the Kriging approach (right). However, besides this micro loop (yellow outline), there also exists a macro loop, which also checks for convergence. Important to realise is, that the micro loop operates on the surrogate model (blue) while the macro loop operates on the real function (red). This conveys, that after the creation of the surrogate model, a procedure similar to RS-RFO is used to find the lowest energy on this predicted surface. Afterwards, it is checked whether this minimum is also the minimum of the real function. If not, the data is taken as an additional sample point, enriching the creation of the next surrogate model. To phrase it differently, the micro iteration loop is used to determine the trial displacement for the macro iteration loop. Since the micro loop operates on the surrogate model, the calculations are very cheap and thus the trial displacement is found very fast.

As conveyed by the yellow outline, the micro loop (right) and RS-RFO approach (left) are very similar. However, minor yet important differences exist. As the actual function of the surrogate model is known, it's Hessian is easily analytically calculated, while this would take a high computational effort when operating on the real function. Thus, the Kriging procedure can make use of the analytical Hessian (light green) of the surrogate model, while RS-RFO uses the guessed Hessian (green). Additionally, as described in 2.3.1, the variance of the surrogate model is known - again in contrast to the real function, where it is not. This allows one to use restricted variance RFO (purple).

A few advantages of this procedure are listed below:

- **Kriging is not a second order Taylor expansion approximation.** Instead, the minimum if the surrogate model is solely dependant on the quality of the prediction. This entails, that the size of the steps taken by the macro loop is not limited or artificially chosen but rather explicitly defined by the surrogate model.

- **Use of restricted variance RFO possible.** The actual function of the surrogate model is known which makes it possible to calculate the variance and use it as a step restriction instead of the step size. This way the procedure has the possibility to take large steps if the variance is low enough and can thus converge more easily.

- **Every sample point adds information.** Therefore the surrogate will always get better with the number of iterations regardless of the step taken. This is in contrast to regular

RS-RFO where a wrongly taken step does not advance the search in any way and can be seen as wasted. This further implies, that with an increasing number of sample points the surrogate model converges towards the real function.

- **Optimization no longer requires an update to the Hessian in each iteration.** Since the Hessian is explicitly described by the surrogate model and analytically calculated in each micro iteration, the normally needed, approximate Hessian update procedure can be completely dropped.

2.3.3 Optimization of hyperparameters θ

As briefly mentioned in section 2.3.1, the used correlation function (equation 13) depends on the so-called hyperparameters, which must be estimated [18]. This can be formally done by maximization of the associated likelihood function. This will be shortly described below.

The likelihood function: Suppose one has gathered a number of data points, as one does in Kriging, and based on a set of hyperparameters, a fit to the data has to be made. Choosing a specific value for each hyperparameter yields a fit, which has a certain probability to accurately predict the unknown function, the data points were originally based upon. Changing the hyperparameters values also changes this probability. The function, describing the probability based upon the variable hyperparameter values, is called the likelihood function. Hence finding its maximum translates into finding the hyperparameters which most likely yield an accurate fit [20].

However, a maximization of said function becomes highly computational expensive for higher-dimensional problems [18], which is true in the here presented case and would thus present a bottleneck for the procedure. Hence, another approach was thought of, using a renowned matrix.

As depicted in figure 2, RS-RFO uses the Hessian to predict the next trial displacement. However, regarding the size of the problem, calculating this exactly would be a huge computational effort. By using the orbital energies, a guessed Hessian can be obtained and utilized for this process. Considering the success of the method, this can thus be seen as a reasonably good approximation of the real Hessian.

It is thus proposed to select the hyperparameters in such a way, that the Hessian produced by the surrogate model exactly recreates this guessed Hessian. Since this Hessian seems be often close to being in a diagonal form, equation 15 [found by coworker Ignacio Fdez. Galván] is suggested.

$$\theta_i = \sqrt{\frac{5E_{ADD}}{3H_{ii}}} \tag{15}$$

This formula is depending on the diagonal values of the guessed Hessian H_{ii} as well as a specific value E_{ADD}, which is part of the trend function (see section 3.3 for further explanation). This formula yields an individual solutions for each hyperparameter θ_i. Masquerading as an ad hoc type of suggestion at first, this can be understood as a pseudo-optimization of the likelihood function. Since the guessed Hessian represents the real Hessian reasonably well, optimizing the hyperparameters such that this Hessian is retained, represents the most likely way of recreating the real function.

However, even though the guessed Hessian is dominantly diagonal, there are significant off-diagonal elements. These are lost in the above mentioned procedure, as only the diagonal elements are considered. Hence finding a way to retain this information would suggest an even better fitting surrogate model.

Thus it is proposed to first diagonalize the guessed Hessian before using it to obtain the hyperparameters. The transformation matrix resulting from this diagonalization is subsequently used to transform displacements and gradients into this new basis before building the surrogate model. After the evaluation of the surrogate model is finished and a set of displacements is chosen, these are transformed back into the old basis and can subsequently be evaluated based on the real function.

Results based on theses different suggestions are presented in section 3.2, whereby the hyperparameters will be called l values from here on out.

3 Results and Discussion

The new procedure has been implemented in the wave function optimization of the open-source OpenMolcas quantum chemistry program package [21].

By using a suitable example, a visual proof of principle is provided in the following section. However, using a larger basis set or a more complicated system rapidly inflates the number of displacement parameters. Hence, beyond the proof of concept, evaluation of the method is done via the iteration counter.
Following the implementation of the new method, several different ways of choosing parameters or optimizing values were thought of. A selection is presented here:

- **Optimization of the l values**
- **Optimization of the trend function**
- **Number of pre-Kriging calculated sample points**

While optimizing another parameter, the following settings were used:

- Utilization of the basis transformation enhanced optimization of l values based on the guessed Hessian
- E_{ADD} was chosen as 1
- Kriging was started after the first iteration

Molecular Structures for benchmarking have been taken from the Baker equilibrium structure test suite [22]. For all benchmarking calculations the Pople type basis set STO-3G [23] was used. Comparisons to the previously reflected conventional methods are made and possible advantages and disadvantages discussed.
A short outlook on remaining challenges is given at the end of the section.

3.1 Proof of concept

In this first section, a short demonstration of the method is given. For this purpose, the wave function of a H_2-molecule, with a fixed geometry, is optimized, using the Dunning type basis set cc-pVDZ [24]. With this setup, there are only two virtual orbitals which are able to mix with one occupied orbital. Thus, only two displacement parameters yield non-zero values, making a graphical presentation possible.

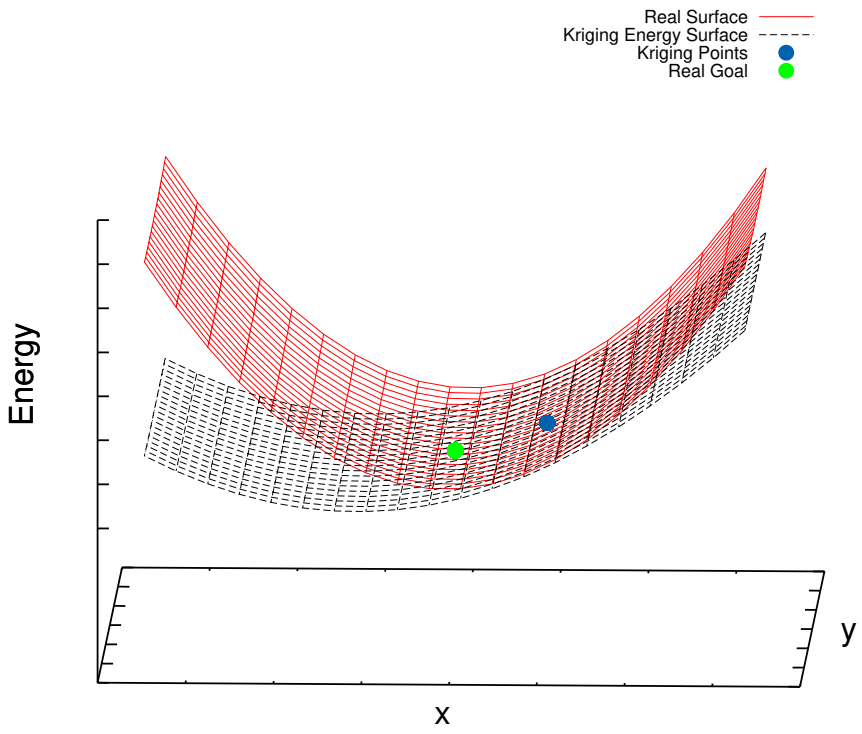
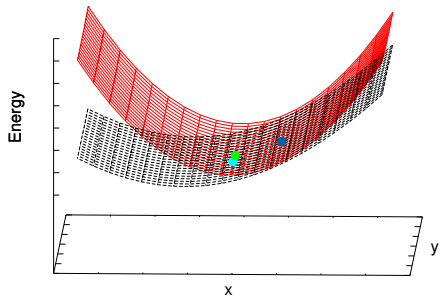
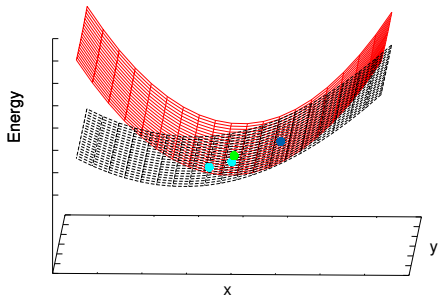


Figure 3: Graphical representation of the surrogate and the real energy surface. The x- and y-axis are the two displacement parameters which have arbitrary values.

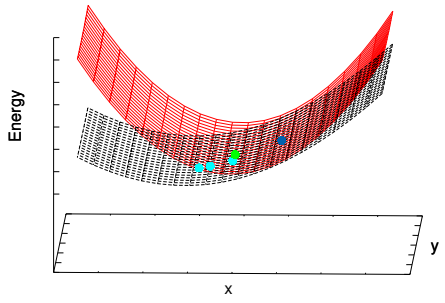
Figure 3 shows the start of the respective optimization via Kriging. The real calculated sample point is given by the blue dot and from its data, the surrogate surface (black) was created to model the real surface. In order to better illustrate the process, the real energy surface (red) as well as the minimum of the same (green) is drawn.



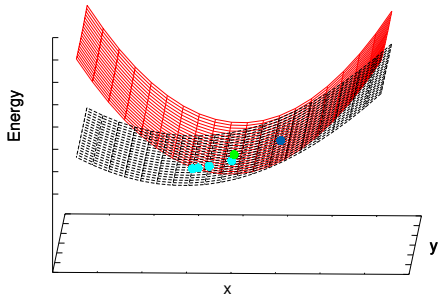
(a) 1. micro iteration



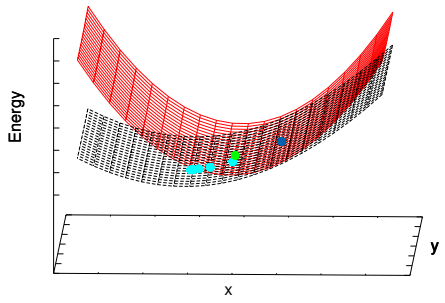
(b) 2. micro iteration



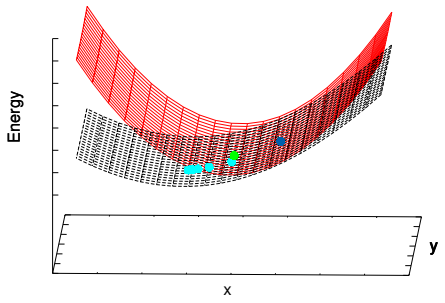
(c) 3. micro iteration



(d) 4. micro iteration



(e) 5. micro iteration



(f) 6. micro iteration

Figure 4: Finding the minimum of the Kriging using RV-RFO. In each picture one more displacement (cyan) is found and evaluated based on the prediction by the surrogate model.

Now that the surrogate model is constructed, RV-RFO is employed to find its minimum as depicted in the micro loop in figure 2. Figure 4 demonstrates this, by showing 6 micro iterations, operating on the surrogate model.

After the micro loop successfully determines the minimum, the resulting displacement parameter is now used as the next step in the macro cycle (see figure 2). This time, the displacement is evaluated based on the real function.

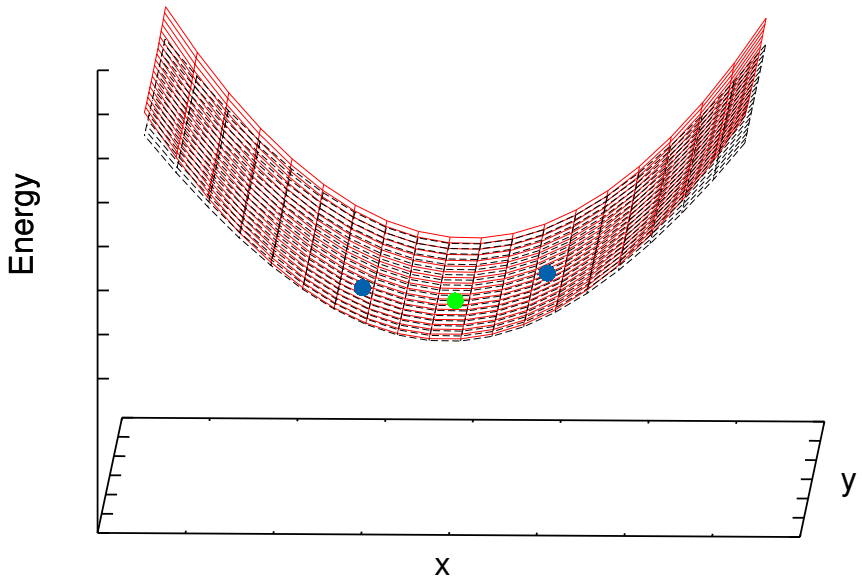


Figure 5: Graphical representation of the surrogate model and the real energy surface with two sample points.

With the information of the additional sample point, a better surrogate model (see figure 5) can be built. Compared to figure 3, the minimum in figure 5 visibly offers a better representation of the real minimum. Continuing this process will thus eventually lead to convergence in the macro cycle within a few iterations, as the surrogate model converges towards the real function.

When analysing figures 3 and 5 one quickly notices that Kriging is most accurate in the immediate vicinity of the sample point(s) and can quickly diverge from the modeled surface beyond that. This is of course because Kriging is by construction an exact interpolation. However, this

doesn't present a problem, since one only wants to recreate a specific part of the function correctly. Having said that, one can notice in figure 5, that the region between sample points is modeled more accurately than beyond. This conveys the idea, that overshooting the real minimum in the beginning can hugely accelerate the calculation, since in that case the real minimum is interpolated rather than extrapolated at the subsequent iteration.

3.2 Determining optimal l values

As discussed in section 2.3.3, the choice of l values is as challenging as it is essential for a working model. Furthermore, equations 13 and 14 dictate the number of l values to be equal to the number of displacement parameters. Hence, avoiding the expensive optimization of all l values by maximization of the associated likelihood function is a non-trivial challenge.

Considering this, an easy and intuitive first approach is to simply set all l values to the same predetermined value, which is empirically optimized. However, even though it is possible to optimize such a value for a specific molecule, it was found to be impossible to determine one, which works for all molecules of the selected suite let alone in general.

Furthermore, it was discovered, that the l values are directly dependant on the trend function mentioned in equation 12, making it impossible, to optimize them on their own. The analytical Hessian of the surrogate model, however, was found to be in turn hugely influenced by the pre-selected l values. This led to the formulation of equation 15 talked about in section 2.3.3.

Implementation of this formula lead to convergence for all molecules in the selected test suite (see table 1) .

Molecule	Kriging	RS-RFO	DIIS
water	8	7	7
ammonia	7	7	7
ethane	7	7	7
acetylene	6	7	7
allene	9	8	8
hydroxylsulphane	9	9	9
benzene	6	6	6
methylamine	8	8	8
ethanol	8	8	9
acetone	11	11	11
disilyl ether	9	8	10
neopentane	7	8	8
furan	10	10	11

Table 1: Iterations needed for convergence to be achieved for different methods.

Results are very comparable to conventional optimization methods. Note, that they very often outperform the DIIS procedure.

As an effort to further improve these results, a procedure was sought to also include the information stored in the off-diagonal elements of the guessed Hessian. Thus the basis transformation procedure mentioned in 2.3.3 was thought of and subsequently tested.

Molecule	Basis transformed Kriging	RS-RFO	DIIS
water	7	7	7
ammonia	7	7	7
ethane	7	7	7
acetylene	7	7	7
allene	8	8	8
hydroxylsulphane	9	9	9
benzene	6	6	6
methylamine	8	8	8
ethanol	8	8	9
acetone	10	11	11
disilyl ether	8	8	10
neopentane	7	8	8
furan	9	10	11

Table 2: Iterations needed for convergence to be achieved for different methods.

Table 2 makes it evident, that this new approach is on par with conventional DIIS as well as RS-RFO in every tested molecule. The number of iterations needed before convergence is reached are either retained or slightly reduced, which suggests a possible replacement of the conventional methods.

3.3 Optimizing the trend function

As explained in section 2.3.1 and seen in equation 12, the trend function has an integral role within the Kriging procedure, which can also be noticed by its in section 3.2 discussed influence on the l value. Since it defines the value of the surrogate model at locations, where the influence of the existing sample points is minimal, it is most influential when few sample points are available. This is especially true as long as the minimum is extrapolated rather than interpolated. Hence, finding a good trend function is crucial for getting the most information in the first iterations.

For the purpose of modeling an energy surface, one has to be certain, that the surrogate model features a minimum, which is non-trivial, especially if there is only one data point available. Thus, to ensure this, the trend function has to be set to a value higher than this calculated point, forcing a minimum somewhere in the vicinity. When adding further sample points, one can additionally make sure, that the trend function always stays above the highest data point. This is a reasonable rule, since the limit of the displacement in infinity should be a higher value than any calculated sample point. As a result, the trend function is set to the value of the highest energy sample point plus a predetermined value defined as E_{ADD} which has to be optimized. This has been done in table 3 by choosing a selection of different addition values.

Molecule	0.01	0.1	1	10	100
water	7	7	7	7	7
ammonia	7	7	7	7	7
ethane	7	7	7	7	7
acetylene	7	7	7	7	7
allene	8	8	8	8	8
hydroxylsulphane	9	9	9	9	9
benzene	6	6	6	6	6
methylamine	8	8	8	8	8
ethanol	8	8	8	8	8
acetone	-	10	10	10	10
disilyl ether	10	8	8	8	8
neopentane	7	7	7	7	7
furan	11	10	9	9	-

Table 3: Iterations needed for convergence to be achieved for different trend functions.

The procedure shows relatively high robustness concerning changes to the trend function. There seems to be no noticeable difference for the smallest molecules while bigger ones can show divergence at extreme values. However, it also defines the margin of acceptable convergences as rather

big. Thus all calculations in this report have been calculated with a value of 1.

3.4 Pre-Kriging calculated sample points

At least one sample point is required in order to construct the surrogate, this is quite intuitive and rational. However, one can also argue, that taking a second or maybe a few more sample points before starting with Kriging can be beneficial, since the initial minima on the surrogate might not provide as much information as sample points found by RS-RFO. Hence it was tested, how many iterations a calculation takes based on how many RS-RFO steps are taken before Kriging is started. The results are depicted in table 4.

Molecule	1	2	3	4	5	6	7	8	9
water	7	7	7	7	7	7	-	-	-
ammonia	7	7	9	9	7	7	-	-	-
ethane	7	7	7	7	7	7	-	-	-
acetylene	7	7	6	7	7	7	-	-	-
allene	8	8	9	8	8	8	8	-	-
hydroxylsulphane	9	9	10	10	9	9	9	9	-
benzene	6	6	6	6	6	-	-	-	-
methylamine	8	8	9	8	8	8	8	-	-
ethanol	8	8	9	10	8	8	8	-	-
acetone	10	10	11	11	11	11	11	11	11
disilyl ether	8	8	9	9	8	8	8	-	-
neopentane	7	7	7	7	8	8	8	-	-
furan	9	10	10	10	10	10	10	10	10

Table 4: Iterations needed for convergence to be achieved for different number of initial sample points.

With only one very specific exception, the number of iterations needed for convergence to be

achieved stay the same or rise for all tested molecules. As a result it can be assumed, that even with as little information as a single data point, the minimum of the Kriging surface offers more information than a RS-RFO step. This can be explained, since RS-RFO will rarely overshoot the minimum and thus the generated sample points still require Kriging to extrapolate.

3.5 Remaining challenges

Even though the presented results seem very promising for Kriging to possible replace the conventional optimizing methods in the future, there are a few remaining challenges which need to be addressed.

- **Overhead added by Kriging increases time consumed in each iteration.** While reducing the amount of iterations needed, Kriging adds overhead to each of them, effectively prolonging the calculation. The existing code sometimes needs as much as 4 times the time for converging in contrast to a conventional calculation which may need an iteration more. This can of course be drastically reduced by addressing the other issues on the list and further code optimisation. But it has to be kept in mind, that a reduction in iteration numbers is worth little when convergence takes more time because of overhead.

- **Procedure requires storing of at least 2 Hessian-sized matrices.** This can quickly lead to memory problems when dealing with a high amount of dimensions [7], which happens quite quickly when calculating larger systems. However, having to store both the guessed Hessian as well as the Kriging Hessian could be avoided by calculating them "on the fly" using the corresponding gradient. This method is used already in conventional RS-RFO and would need some additional coding to be implemented into the new routine.

- **Optimization of l values requires diagonalization of the Hessian.** As explained in section 2.3.3, this needs to be done in order to perfectly replicate the guessed Hessian when operating on the surrogate model. Additionally one thus has to store the transformation matrix, converting the displacements and gradients of the surrogate into the new basis of the guessed Hessians eigenvectors. As a result, there is ultimately a need for an easier and quicker way of optimising the l values, while retaining an equal level of accuracy.

- **Empirical optimization of the trend function probably not universally optimal.** The number of iterations needed is hugely dependant on the amount of information gathered in the first sample points. That is, as long as the minimum is extrapolated rather than interpolated by the model, the outcome is hugely dependant on the trend function. Hence a possible goal could be to find a way to optimize the trend function non-empirically such

that the surrogate overshoots the minimum in a favorable way as soon as possible. Thus interpolation could start earlier and the procedure would converge quicker.

- **High dimensionality of the covariance matrix.** As it is now, the creation of the surrogate model consumes the lion's share of the time spend on the whole procedure. This can be accounted to the enormous size of the covariance matrix, which is a part of the stationary covariance process $\delta(\mathbf{x})$ mentioned in equation 12. It currently has approximately the size of the hessian times the square number of sample points. This is a major contrast to the application of Kriging in geometry optimization, since here one has way more than 3 dimensions for each atom. Thus, this increasing size is much sooner, i.e. with much smaller molecules, a problem in wave function optimization than it is in geometry optimization. Exploring options to reduce the dimensionality by omitting non-essential information could be the key to solve this issue, albeit a certain overhead will remain.

4 Outlook

The results presented in this report make a strong argument for using a surrogate model approach to find the optimized orbital coefficients. While still being in the stage of development with many code snips still being improvable, the routine is seemingly on par with conventional calculations using DIIS and RS-RFO regarding their iteration requirements. Considering the success of the method in geometry optimization reported by Raggi *et al.* [25], it is fair to assume that the challenges listed in section 3.5 can be partially addressed or met completely.

Addressing one of the noted problems, finding a way to retain the present level of accuracy without having to diagonalize the guessed Hessian would be vital for calculations of bigger molecules and thus presents a worthwhile target for future improvement.

Additionally, as mentioned in section 3.1, Kriging is by construction an exact interpolation, which in turn suggests that one might want to optimize the trend function such, that it overshoots the target in the first iteration. Denzel *et al.* [5] suggest such a procedure in a geometry optimization application of Kriging, intentionally overshooting and thus reaching convergence earlier. Implementing such a technique in wave function optimization could therefore further enhance the method's merits compared to conventional procedures.

References

[1] Alán Aspuru-Guzik, Roland Lindh, and Markus Reiher. The matter simulation (r) evolution. *ACS central science*, 4(2):144–152, 2018.

[2] Thomas H Fischer and Jan Almlof. General methods for geometry and wave function optimization. *The Journal of Physical Chemistry*, 96(24):9768–9774, 1992.

[3] Ron Shepard, Isaiah Shavitt, and Jack Simons. Comparison of the convergence characteristics of some iterative wave function optimization methods. *The Journal of Chemical Physics*, 76(1):543–557, 1982.

[4] Selvakumar Ulaganathan, Ivo Couckuyt, Tom Dhaene, Joris Degroote, and Eric Laermans. Performance study of gradient-enhanced kriging. *Engineering with computers*, 32(1):15–34, 2016.

[5] Alexander Denzel and Johannes Kästner. Gaussian process regression for geometry optimization. *The Journal of Chemical Physics*, 148(9):094114, 2018.

[6] Alexander Denzel and Johannes Kastner. Gaussian process regression for transition state search. *Journal of Chemical Theory and Computation*, 14(11):5777–5786, 2018.

[7] Péter Pulay. Convergence acceleration of iterative sequences. the case of scf iteration. *Chemical Physics Letters*, 73(2):393–398, 1980.

[8] A Banerjee and F Grein. Convergence behavior of some multiconfiguration methods. *International Journal of Quantum Chemistry*, 10(1):123–134, 1976.

[9] Emili Besalú and Josep Maria Bofill. On the automatic restricted-step rational-function-optimization method. *Theoretical Chemistry Accounts*, 100(5-6):265–274, 1998.

[10] Konstantin N Kudin, Gustavo E Scuseria, and Eric Cances. A black-box self-consistent field convergence algorithm: One step closer. *The Journal of chemical physics*, 116(19):8255–8261, 2002.

[11] Ron Shepard and Michael Minkoff. Some comments on the diis method. *Molecular Physics*, 105(19-22):2839–2848, 2007.

[12] Thorsten Rohwedder and Reinhold Schneider. An analysis for the diis acceleration method used in quantum chemistry calculations. *Journal of mathematical chemistry*, 49(9):1889, 2011.

[13] Gustavo E Scuseria, Timothy J Lee, and Henry F Schaefer III. Accelerating the convergence of the coupled-cluster approach: The use of the diis method. *Chemical physics letters*, 130(3):236–239, 1986.

[14] G Berthier, M Defranceschi, and C Le Bris. Shortcomings in computational chemistry. *International journal of quantum chemistry*, 93(3):156–165, 2003.

[15] Luc Laurent, Rodolphe Le Riche, Bruno Soulier, and Pierre-Alain Boucard. An overview of gradient-enhanced metamodels with applications. *Archives of Computational Methods in Engineering*, 26(1):61–106, Jan 2019.

[16] Ajit Banerjee, Noah Adams, Jack Simons, and Ron Shepard. Search for stationary points on surfaces. *The Journal of Physical Chemistry*, 89(1):52–57, 1985.

[17] Jack PC Kleijnen. Kriging metamodeling in simulation: A review. *European journal of operational research*, 192(3):707–716, 2009.

[18] Mohamed A Bouhlel and Joaquim RRA Martins. Gradient-enhanced kriging for high-dimensional problems. *Engineering with Computers*, 35(1):157–173, 2019.

[19] Peter Guttorp and Tilmann Gneiting. Studies in the history of probability and statistics xlix on the matern correlation family. *Biometrika*, 93(4):989–995, 2006.

[20] In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100, 2003.

[21] Ignacio Fdez Galván, Morgane Vacher, Ali Alavi, Celestino Angeli, Francesco Aquilante, Jochen Autschbach, Jie J Bao, Sergey I Bokarev, Nikolay A Bogdanov, Rebecca K Carlson, et al. Openmolcas: From source code to insight. *Journal of Chemical Theory and Computation*, 2019.

[22] Jon Baker. Techniques for geometry optimization: A comparison of cartesian and natural internal coordinates. *Journal of computational chemistry*, 14(9):1085–1100, 1993.

[23] Warren J Hehre, Robert F Stewart, and John A Pople. self-consistent molecular-orbital methods. i. use of gaussian expansions of slater-type atomic orbitals. *The Journal of Chemical Physics*, 51(6):2657–2664, 1969.

[24] Thom H Dunning Jr. Gaussian basis sets for use in correlated molecular calculations. i. the atoms boron through neon and hydrogen. *The Journal of chemical physics*, 90(2):1007–1023, 1989.

[25] Gerardo Raggi, Christian L. Ritterhoff, Ignacio Fdez. Galván, Morgane Vacher, and Roland Lindh. Yet unpublished manuscript.