Parameter reconstructions are indispensable in metrology. Here, the objective is to explain $K$ experimental measurements by fitting to them a parameterized model of the measurement process. The model parameters are regularly determined by least-square methods, that is, by minimizing the sum of the squared residuals between the $K$ model predictions and the $K$ experimental observations, $\chi^2$. The model functions often involve computationally demanding numerical simulations. Bayesian optimization methods are specifically suited for minimizing expensive model functions. However, in contrast to least-square methods such as the Levenberg–Marquardt algorithm, they only take the value of $\chi^2$ into account, and neglect the $K$ individual model outputs. A Bayesian target-vector optimization scheme with improved performance over previous developments, that considers all $K$ contributions of the model function and that is specifically suited for parameter reconstruction problems which are often based on hundreds of observations is presented. Its performance is compared to established methods for an optical metrology reconstruction problem and two synthetic least-squares problems. The proposed method outperforms established optimization methods. It also enables to determine accurate uncertainty estimates with very few observations of the actual model function by using Markov chain Monte Carlo sampling on a trained surrogate model.

1. Introduction

A common task in science and engineering is the fitting of the parameters of a model function in order to match the outputs of the model to an experimental observation. Often, one has to match $K$ experimental observations simultaneously, which can be done by minimizing the sum of the squared residuals $\chi^2$ between model outputs and experimental observations. Generally, this is a non-linear least-squares problem, and is regularly solved by iterative numerical schemes, for example, by the Gauss–Newton method or by the Levenberg–Marquardt algorithm.\[1–3\]

The models involved in these inverse problems — for example, in optical metrology\[4,5\] — are often very costly to evaluate, because they frequently involve the solution of differential equations via numerical simulations. In scatterometry for example one solves Maxwell’s equations to simulate the scattering process of light off a nanostructured sample.\[6\] This means that depending on the complexity and size of the numerical model, a single evaluation can take several minutes or more to complete. This can present a problem, as many optimization schemes do not handle expensive model functions in an efficient way. Typically, they calculate the parameters of the next iteration based on only a few previous model evaluations, neglecting the information from other evaluations of the optimization history. An additional issue for many model functions is the difficulty of obtaining their derivatives with respect to the model parameters. Optimization schemes such as the Levenberg–Marquardt algorithm require derivatives in order to generate a sampling candidate for the next iteration. Frequently, derivatives are generated by means of a finite differences scheme, which is inefficient and often inaccurate. Therefore, parameter reconstructions in metrology applications can place high demands on computing resources and time. This creates the need for efficient and robust reconstruction methods that make optimal use of energy and time consuming model evaluations.

Bayesian optimization (BO)\[7,8\] methods are sequential optimization methods that satisfy the resource requirement, and are therefore an appropriate choice when optimizing expensive black-box model functions. At every BO iteration a stochastic model function—most often a Gaussian process (GP)\[9\]—is trained using all previous observations of the model function. This stochastic model is then used to find input parameters for the next iteration, which, for example, lead to a large expected improvement over the currently known minimum.\[10\]
Recent studies\cite{12,13} considered the extension of the BO approach for iteratively solving least-squares problems by training $K$ GPs related to each of the $K$ data channels of the model function. In these works, the approach was successfully applied for problems with $K \leq 20$ channels. Huang et al.\cite{13} applied these approaches to calibrate the parameters $p \in \mathbb{R}^K$ of an expensive uni-variate scalar function $f_i(t)$ to measurements for thousands of times $t_i$. In order to reduce the dimensionality of the measurements and the corresponding functional response to only $K = 3$, a functional principal component analysis was applied, based on $10^4$ space-filling parameter samples prior to solving the least-square problem. In another study\cite{14} a multi-output GP\cite{15,16} was used in a non-iterative parameter reconstruction approach. That is, the GP was trained with a precomputed set of simulation results to directly infer parameter values and uncertainties. This led to higher reconstruction accuracies than a library lookup method.

In this study we start from the approach taken by Uhrenholt and Jensen\cite{12} to develop an algorithm for iterative parameter reconstruction problems, which often have a large number of data channels $K$, for example, a few hundred.\cite{17–19} We find that this creates two particular challenges. First, BO methods are typically associated with a computational overhead of a few seconds when determining new parameter candidates to sample, which is caused by the training of the surrogate model. Training and evaluating $K \approx 100$ independent Gaussian processes dramatically increases this overhead. We propose to use a shared covariance structure to largely limit the additional overhead. In doing this we assume implicitly that each data channel has a comparable structure to largely limit the additional overhead. We propose to use a shared covariance structure to largely limit the additional overhead. In doing this we assume implicitly that each data channel has a comparable structure to largely limit the additional overhead. In doing this we assume implicitly that each data channel has a comparable structure to largely limit the additional overhead. In doing this we assume implicitly that each data channel has a comparable structure to largely limit the additional overhead.

Second, in order to be fast, the scheme approximates the probability distribution of $\chi^2$ as an ordinary non-central chi-squared distribution with $K$ degrees of freedom (DoFs). This appears to work well for a small number of data channels, as is evident from the results presented in ref.\cite{12}. However, for a large number of DoFs $K$ we find that this leads to a very inefficient optimization scheme, as the probability distribution largely underestimates the probability of finding small $\chi^2$ values. We find that using an effective number of DoFs $\tilde{K}$, which is often much smaller than $K$, leads to good optimization performance. We propose to choose the value of $\tilde{K}$ with maximum-likelihood of the approximate chi-squared probability distribution for all previous $M$ observations of the model function.

Compared with other approaches, we show that the proposed Bayesian target-vector optimization (BTVO) scheme often requires significantly fewer iterations to reconstruct the desired model parameters. Another important advantage of the approach is that each GP offers a good non-linear model of the corresponding data channel. This enables, for example, to quickly sample from the approximated posterior probability distribution of the model parameters and to accurately quantify parameter uncertainties.

The paper is organized as follows. In Section 2 we give a theoretical introduction into parameter reconstructions and into solving the corresponding least-squares problems. We review BO methods and their extension by Uhrenholt and Jensen. We then discuss the shortcomings for a large number of data channels $K$ and introduce a mitigation strategy. We then discuss Markov chain Monte Carlo (MCMC) sampling for determining the uncertainties of the reconstructed parameters, and we describe a way to drastically reduce the required number of model evaluations by training and evaluating surrogate models. In Section 3 we compare the performance of the proposed BTVO against a selection of established optimization methods. We consider three different model functions: a computationally intensive real-world optical metrology example, and two analytical model functions obtained from the NIST Standard Reference Database.\cite{20} The two analytical model functions are additionally used to highlight the benefit of derivative information for the reconstruction performance. In Section 4 we employ one of the analytical model functions further to demonstrate the efficiency and accuracy of the surrogate model augmented MCMC method. Finally, we use the surrogate augmented MCMC to discuss correlations in the experimental problem discussed in ref.\cite{21}.

2. Theoretical Background

Parameter reconstructions are often based on fitting the vectorial output of a parameterized model function $f(p)$ to an experimental measurement $t = (t_1, \ldots, t_K)^T$, where $p \in \mathcal{X} \subset \mathbb{R}^n$ and $f : \mathcal{X} \rightarrow \mathbb{R}^K$. The model function is treated as a black-box that can be evaluated point-wise. It is assumed that $f(p)$ is once differentiable and that the $K$ individual channels can be modeled by Gaussian processes. As usual, we further assume that the model describes the measurement process sufficiently well, such that model errors can be neglected. Measurement noise is modeled by assuming that the $i$th measurement value is equal to the model value for the true parameter $p$, plus some noise contribution,

$$t_i = f_i(p_i) + \varepsilon_i$$  \hfill (1)

Usually, the noise is modeled to be normally distributed with zero mean and variance $\eta_i^2$, that is, $\varepsilon_i \sim \mathcal{N}(0, \eta_i^2)$.

Finding a good estimate for $p_i$ can be considered an optimization task. We can obtain the least-square estimate (LSQE)

$$p_{\text{LSQE}} = \arg \min_{p \in \mathcal{X}} \chi^2(p)$$  \hfill (2)

by minimizing the sum of the squared residuals

$$\chi^2(p) = \sum_{i=1}^{K} \frac{(f_i(p) - t_i)^2}{\eta_i^2}$$  \hfill (3)

Having determined $p_{\text{LSQE}}$, one is often also interested in the local probability distribution of the parameter values in order to determine confidence intervals of the reconstructed parameter values (c.f. Section S1, Supporting Information).

In optical metrology applications, evaluating the model function $f$ typically involves running a numerical simulation, which can include the assembly of the discretized problem, numerical solution of differential equations, and postprocessing of the results. Calculating the result of the model function for a single set of parameters can therefore take a lot of computation time, depending on the complexity of the model and on numerical accuracy requirements.
2.1. Established Approaches for Parameter Reconstruction

Minimizing Equation 3 can be done, for example, using local methods such as Nelder-Mead or the Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS), or using global heuristic methods, for example particle swarm optimization,[22,23] or differential evolution,[24,25] or by maximizing the appropriate likelihood function using MCMC sampling methods.[26] Since minimizing Equation 3 is a least-square problem, it can of course also be solved using least-square methods like the Gauss–Newton scheme or the Levenberg–Marquardt algorithm,[17] which directly minimizes the residuals between model output and experimental results.

An estimate for the confidence intervals can be obtained by exploiting information about the derivatives of the model function with respect to each parameter at the point estimate \( p_{LSQE} \). As is for example available after minimizing Equation 3 using least-squares methods,[27–29] or by applying MCMC sampling methods to the appropriate likelihood function.[30] The latter has the advantage that one can obtain accurate uncertainties in terms of 16%, 50% (i.e., the median), and 84% percentiles of the actual model parameter distributions, as well as determine non-linear correlations between model parameters. Methods that exploit local derivative information usually only yield approximate Gaussian parameter uncertainties in terms of \( \sigma \) intervals, and are only capable of establishing linear correlations between model parameters. We denote these Gaussian parameter uncertainties as \( \epsilon_{LSQE} \).

Using these established approaches to perform parameter reconstructions can be a computationally costly endeavor. Standard least-squares algorithms reconstruct the first minimum that they reach, which depends on the initial guess fed into the method. In more complex energy landscapes this can be a local minimum rather than the point estimate \( p_{LSQE} \) one seeks, a problem which can be alleviated, for example, by using a multi-start approach. This does however not guarantee that \( p_{LSQE} \) is found. Particle swarm optimization and differential evolution on the other hand are not designed to be efficient in the sense that they obtain the point estimate in as few evaluations of the model function as possible, and MCMC sampling methods even rely on the fact that they evaluate the model function very often, since the quality of the reconstruction increases with the number of samples drawn.[11] This can present an issue from a resource standpoint, and necessitates more resource efficient optimization methods.

2.2. Bayesian Optimization Approaches for Parameter Reconstruction

Bayesian optimization (BO) methods[7,8] are sequential optimization methods and are known for being very efficient at performing global optimizations of expensive black-box functions.[10,32] To this end BO methods train a stochastic surrogate model—most often a Gaussian process (GP)[9]—in an iterative fashion, using all previous observations of the model function after \( m \) iterations. This surrogate model is usually much quicker to evaluate than the model function itself. The predictions made by the surrogate model are then used by an acquisition function to determine a parameter \( p_{m+1} \) which is beneficial to sample the model function with next. The exact meaning of “beneficial” in this context depends very much on the strategy pursued by the acquisition function, as well as the optimization goal.

In Section 2.2.1 we shortly introduce GPs and Gaussian process regression and describe in Section 2.2.2 how it is used in “conventional” BO for minimizing \( f(p) \). In Section 2.2.3 we will discuss the BTTO approach proposed by Uhrenholt and Jensen,[12] which minimizes Equation 3 by considering the individual contributions of the components of \( f(p) \).

2.2.1. Gaussian Process Regression

GPs are stochastic processes that are defined on a continuous domain \( X \subset \mathbb{R}^N \). A random function \( f \) is a GP, if for any finite tuple \( X = [p_1, \ldots, p_N] \in \mathbb{R}^N \) the random vector \( Y = [f(p_1), \ldots, f(p_N)]^T \) is a multivariate normal distribution (MVN). As such, GPs are an extension of finite dimensional MVNs to an infinite dimensional case.[13] A GP is completely specified by a mean function \( \mu : X \rightarrow \mathbb{R} \) (which replaces a mean vector in the finite dimensional case) and a covariance kernel function \( k : X \times X \rightarrow \mathbb{R} \) (replacing a covariance matrix of a finite dimensional MVN).[9] A GP can be trained to determine the posterior distribution of function values given some observations of the function, and can then serve as a stochastic predictor or interpolator for the training data. A usual choice for the prior mean and covariance kernel function, which is also considered in the following, are a constant mean function and the Matérn 5/2 covariance function[34]

\[
\mu(p) = \mu_0 \\
k(p, p') = \sigma_0^2 \left( 1 + \sqrt{5r} + \frac{5}{3}r^2 \right) \exp \left(-\sqrt{5r}\right)
\]

where \( r = \sqrt{\sum_{i=1}^{N} \frac{(p_i - p_i')^2}{l_i}} \) (4)

Up to some maximum number of observations \( M_{\text{hyper}} \) the hyper-parameters \( \mu_0, \sigma_0, l_1, \ldots, l_K \) are chosen to maximize the likelihood of the observations.[15,36] For this, the length scales \( l_1, \ldots, l_K \), which enter the covariance function in a non-trivial way, are kept constant. A GP trained on function evaluations \( Y = [f(p_1), \ldots, f(p_M)]^T \) allows to make predictions for any parameter vector \( p^* \) in the form of a normally distributed random variable \( \tilde{f}(p^*) \sim \mathcal{N}(\tilde{\mu}(p^*), \sigma^2(p^*)) \) with mean and variance

\[
\tilde{\mu}(p^*) = \mu_0 + k^T(p^*)K^{-1}[Y - \mu_0] \\
\sigma^2(p^*) = \sigma_0^2 - k^T(p^*)K^{-1}k(p^*)
\]

where \( k(p^*) = [k(p^*, p_1), \ldots, k(p^*, p_M)]^T \) and \( (K)^{ij} = k(p_i, p_j) \).

For better numerical stability, the positive semidefinite covariance matrix \( K \) is not inverted directly. Instead, one can compute its Cholesky decomposition \( K = L K^T \) into a lower and upper triangular matrix in \( O(M^3) \) steps for \( M \leq M_{\text{hyper}} \). For constant length scales (i.e., \( M > M_{\text{hyper}} \)) an update of the decomposition only requires \( O(M^2) \) steps.[36] Afterward, one solves

\[
L^T L^{-1} \alpha = Y - \mu_0 \mathbf{1}
\]
for $\alpha$ by forward and backward substitution and

$$L \alpha = \beta (p^*) = k(p^*)$$

(8)

for $\beta(p^*)$ by forward substitution both in $\mathcal{O}(M^2)$ steps. With the auxiliary vectors $\alpha$ and $\beta$, Equations 5 and 6 can be evaluated as $y(p^*) = \mu_k + k(p^*)^T \alpha$ and $\sigma^2(p^*) = \sigma^2_k - \beta^T \beta$ in only $\mathcal{O}(M)$ steps. For the computation of the next sampling point, one requires predictions for many points $p^*$. Hence, solving Equation 8 for $A$ different values of $p^*$ (typically $A \geq 1000$) requires a large fraction of the computation time.

Training of GPs can be easily extended to exploit derivative information if available. When solving Maxwell’s equations this can be calculated for example, using the direct method or the adjoint method.\(^{[36]}\) While this enlarges the number of data points and the size of the covariance matrix accordingly, all the above considerations are equally valid.

### 2.2.2. Bayesian Optimization

First, we consider the conventional BO method in order to minimize the scalar function $\chi^2(p)$ defined in Equation 3. At each iteration $m$, the BO approach employs the predictions made by the trained GP in order to determine the next sampling point $p_{m+1}$. This point is selected according to some infill criterion at the maximum of an acquisition function $\alpha(p)$.\(^{[7,8]}\) A usual choice is the expected improvement (EI) with respect to the lowest known function value $\chi^2_{\text{sup}} = \min\{\chi^2(p_1), \ldots, \chi^2(p_i)\}$.

The corresponding acquisition function is defined as

$$\alpha_{\text{EI}}(p) = E \left[ \min(0, \chi^2_{\text{min}} - f(p)) \right]$$

(9)

where $f(p)$ is a Gaussian random variable with mean $\overline{f}(p)$ and variance $\sigma^2(p)$ given in Equations 5 and 6. Another infill criterion is the lower confidence bound (LCB) with acquisition function

$$\alpha_{\text{LCB}}(p) = \kappa \sigma^2(p) - \overline{f}(p)$$

(10)

where $\kappa$ is a scaling factor.

### 2.2.3. Bayesian Target-Vector Optimization

Directly minimizing the squared error function Equation 3 using BO (cf. Figure 1) has some important drawbacks. First, it ignores the knowledge of the function values $f_i(p_1), \ldots, f_i(p)$ that contribute to the value of $\chi^2(p)$ in Equation 3. This information loss leads to a significantly slower convergence. The second issue stems from the fact that a GP can only be used to predict normal distributions, that is, Equations 5 and 6. The predictions of a GP that has been conditioned on observations of Equation 3 are necessarily incorrect, since $\chi^2(p)$ does not follow a normal distribution, but rather a chi-squared distribution. This problem becomes especially pronounced when the GP predicts small mean values and large variances, as this may lead the acquisition function into exploring regions where the predictions suggest an improvement to negative $\chi^2$ mean values. This clearly conflicts with $\chi^2$ being larger than, or equal to, zero. To address the issues, we follow the approach proposed by Uhrenholt and Jensen,\(^{[12]}\) in which each of the $K$ components of $f(p)$ is modeled by a GP.

Using $K$ independent GPs for making a regression on $K$ channels increases the computational effort to $\mathcal{O}(K \cdot M^2)$ steps for computing $K$ Cholesky decompositions. For $M > M_{\text{hyper}}$ observations we keep the length scales constant and the computational effort increases to $\mathcal{O}(A \cdot K \cdot M^2)$ steps for making $A$ different predictions in order to maximize the acquisition function. For metrology applications with often more than 100 channels, the corresponding computation times render the approach impractical. Therefore, in contrast to the approach of Uhrenholt and Jensen, we propose to model the GPs using the same covariance kernel function and only allow for different hyperparameters $\mu_{\text{c}}^{(i)}$ and $\sigma_{\text{c}}^{(i)}$ for each channel $i = 1, \ldots, K$. The optimal value of the hyperparameters $\mu_{\text{c}}^{(i)}, \sigma_{\text{c}}^{(i)}$ for $i = 1, \ldots, K$ and the length scale parameters $l_1, \ldots, l_N$ for the shared covariance kernel matrix are chosen by maximizing the likelihood of the training data averaged over all channels. Hence, the Cholesky decomposition has to be computed only once and can be used for all channels as well as for the solution of Equation 8. Only Equation 7 has to be solved for all $K$ channels since it depends on the acquired function values $Y$ in each channel. However, the equation has to be solved only once to make an arbitrary number $A$ of predictions for different parameter vectors $p^*$. Since $A$ is usually much larger than $K$, the computational overhead of $\mathcal{O}(K \cdot M^2)$ steps for solving Equation 7 for each channel turns out to be acceptable in comparison to solving Equation 8 repeatedly in $\mathcal{O}(A \cdot M^2)$ steps.

The $K$ GPs each provide predictions in form of a normally distributed random variable $\hat{f}_i(p) \sim N(\overline{f}_i(p), \sigma_i^2(p))$, for $i = 1, \ldots, K$, such that the prediction of $\chi^2(p)$ defined in Equation 3 is a random variable

$$\chi^2(p) = \sum_{i=1}^{K} \frac{(\hat{f}_i(p) - f_i)^2}{\eta_i^2}$$

(11)
In order to demonstrate the issues that arise for a large number of data channels for this series of approximations, we consider a Wilson–Hilferty approximation of the non-central chi-squared distribution, according to which the random variable

$$z(p) = \left( \frac{\gamma^{-2}(p)}{K + \lambda(p)} \right)^{\frac{1}{2}}$$

(14)

approximately follows a normal distribution with mean \( \bar{z}(p) = 1 - \frac{1}{\sqrt{K}} \) and variance \( \sigma(z)^2(p) \equiv \frac{1}{K} \), where \( \nu(p) \equiv (K \pm p) / K + 2 + \frac{p}{K} \).

We consider the limit of \( K \to \infty \) while keeping the non-centrality \( \lambda(p) / K \) per channel constant. This is a good approximation for the case that the density of data channels in an experiment, for example, the density of a measured spectrum, is increased. When increasing \( K \), we observe that \( \sigma(z)^2(p) \to 0 \), which means that the PDF approaches a \( \delta \)-distribution and the CDF a step function. Since \( \bar{z}(p) \to 1 \), the maximum of the PDF approaches

$$\chi^2_{mode} = \frac{\gamma^{-2}(p)}{K + \lambda(p)} = \sum_{i=1}^{K} \frac{(\hat{y}_i(p) - t_i)^2}{\sigma_i^2(p)}$$

(15)

Let \( p_{\min} \) be the parameter with the minimal observed value of \( \chi^2 \). Moving away from \( p_{\min} \) to regions with fewer training samples, \( \bar{y}(p) \) approaches the mean value of the corresponding GP, \( \mu(p) \), which has generally a larger deviation from the measurement \( t \). Moreover the GP uncertainty increases. Hence, using this approximated probability function for large \( K \) the probability density of seeing \( \chi^2 \) values smaller or close to \( \chi^2(p_{\min}) \) quickly approaches zero for points not close to \( p_{\min} \). Consequently, the employed infill criterion can only select points close to \( p_{\min} \). In contrast to conventional BO, this leads to a very localized optimization behavior with possibly slow convergence and no exploration of regions with fewer training data. The selection of sample candidates that are too close to previously sampled positions also has a negative impact on the numerical stability of the scheme, as it leads to an ill-conditioned covariance matrix \( K \). Therefore, the optimization is terminated if the infill criterion is only able to select samples that are within a distance of \( 1 \times 10^{-3} \) length scales to previously sampled positions.

This undesired behavior of the approximate probability distribution for large \( K \) is a consequence of modeling all channels by independent random variables. For parameter reconstruction problems we know, however, that there exist parameter values for which all \( K \) residuals are small at the same time—an event that is expected to be almost impossible if \( K \) is large. Moreover, channels that belong to similar experimental conditions (e.g., similar angles and wavelengths in scatterometry experiments) show a similar dependence on the parameters \( p \) and are thus positively correlated. If sum rules apply to the measurement process (e.g., energy or momentum conservation) also negative correlations between the channels can be observed. As described above, neglecting these \( p \)-dependent correlations is done for performance reasons.

To mitigate these issues, we propose to replace the number of DoFs \( K \) in the parameterization of the approximated probability distribution by an effective number of DoFs \( \tilde{K} \), such that the approximate probability distribution better matches all
M previous observations $Y$ of the model function. To this end, we take the sum over all $\chi^2_p$ values, $\chi^2_p = \sum_{i}^{M} \chi^2_{oi}$, find the corresponding approximate marginal likelihood distribution, and then maximize the likelihood of the observed $\chi^2_p$ value with respect to the effective DoFs to obtain the maximum-likelihood estimate $K_{\text{MLE}}$. The acquisition function is then evaluated based on a non-central chi-squared distribution with $K = K_{\text{MLE}}$ effective DoFs and non-centrality defined in Equation 13. The derivation of $K$ is detailed in Section S3, Supporting Information. We find that this yields an effective way to regain the exploratory nature of the BO method also for the BTVO with a large number of data channels $K$. A complete algorithmic overview of the BTVO scheme can be found in Algorithm 1 in the Supporting Information.

### 2.3. Markov Chain Monte Carlo for Accurate Parameter Uncertainties

By exploiting information from the Jacobian of the model function at the LSQE $p_{\text{LSQE}}$ it is possible to give estimates for the parameter uncertainties $\epsilon_{\text{LSQE}}$, c.f. also Section S2, Supporting Information. These uncertainty estimates are generally symmetrical, which does not necessarily reflect on the true model parameter distribution. Furthermore, it is only possible to obtain approximate linear correlations between model parameters, while one is often interested in more complex model parameter relationships when performing parameter reconstructions.\[^{[30]}\]

More accurate parameter uncertainties beyond the Gaussian approximation can be given in terms of 16%, 50%, and 84% percentiles of the likelihood $L(p)$ or their posterior probability $P(p)$ of the parameter values (c.f. Section S1, Supporting Information). The 50% percentile, that is the median, then represents the maximum likelihood estimate (MLE) (when sampling the likelihood function) or the maximum a-posteriori estimate (MAP) (when sampling the posterior density).

Unfortunately, for the non-Gaussian probability distributions $L(p)$ and $P(p)$, the percentiles can in general not be determined analytically. Instead, one determines them based on a large set of samples whose density in the parameter space is proportional to the probability distribution of interest. These sets are typically generated by Markov chain Monte Carlo (MCMC) sampling methods.\[^{[31,41,42]}\] The sampling sets can be also used to expose the Gaussian prediction $f(p)$ of the model function, which is then used to generate the equilibrium distribution. We have to assert that the surrogate model is an accurate representation of the model function in the region of interest, which is often located around the acquired LSQE point. Because the surrogate model was created during a parameter reconstruction, this region is often only explored reasonably well in the direction from which the LSQE was found. To reduce the uncertainty of the surrogate model in the rest of the region of interest, we enter a refinement stage in which we actively train the model with more parameter samples close to the LSQE point $p_{\text{LSQE}}$. To this end, we draw $S$ random samples $p_1, \ldots, p_S$ from the multivariate normal distribution $\mathcal{N}(p_{\text{LSQE}}, \text{Cov}(p_{\text{LSQE}}))$ defined by the parameter covariance matrix $\text{Cov}(p_{\text{LSQE}})$ at the LSQE point (c.f. Section S2, Supporting Information) and evaluate the forward model at the point with maximum mean uncertainty of the surrogate models of all channels, that is,

$$p_{\text{max}} = \arg\max_{p \in \mathcal{P}} \frac{1}{K} \sum_{k=1}^{K} \sigma_k(p)$$

In essence, we follow a sequential experimental design strategy in which the criterion is the predicted variance of the surrogate model.\[^{[46]}\] We stop the additional sampling of the forward model when the maximum mean uncertainty is below some threshold $\sigma_{\text{min}}$. An algorithmic overview can be found in Algorithm 2 in the Supporting Information.

### 3. Benchmarks

In order to assess the performance of the proposed BTVO scheme, we applied it to three parameter reconstruction problems: an experimental dataset, where the model function is a finite element simulation, and two synthetic datasets with analytic model functions.

The experimental dataset has been measured during a grazing incidence X-ray fluorescence (GIXRF) experiment at the synchrotron radiation source BESSY in Berlin.\[^{[21]}\] The analytic datasets were obtained from the NIST Standard Reference Database\[^{[40]}\] for non-linear regression problems. To show resilience of the reconstruction algorithm, we chose two datasets that differ with respect to the number of free parameters and data points: the MGH17 dataset\[^{[47]}\] contains five free parameters and 33 data points, while the Gauss3 dataset\[^{[48]}\] contains eight free parameters and 250 data points. The Gauss3 dataset is of particular interest, as it is comparable to problems from optical metrology in terms of free parameters and data points. Because of their analytical nature, derivatives with respect to all free parameters are easily calculated.

The proposed BTVO method was compared in a benchmark type analysis, where its reconstruction performance was...
compared to that of other optimization schemes. Here, Levenberg–Marquardt (LM)[1,2,49] (also c.f. Section S2, Supporting Information, for the benchmarks we employed the scipy implementation `scipy.optimize.least_squares’[50], BO as detailed in Section 2.2.2, the limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm with box constraints (L-BFGS-B),[51] as well as the Nelder–Mead (NM) downhill simplex algorithm[52] were used.

Of these methods, only BTVO and LM are methods that can solve least-squares problems natively by utilizing all data channels of the model function. BO, L-BFGS-B, and NM are optimization methods that are usually used for the minimization or maximization of scalar functions. These methods therefore minimize (functions of) $\chi^2(p)$ as defined in Equation 3 directly. L-BFGS-B and NM both took the regular scalarized value $\chi^2(p)$ as model function, where L-BFGS-B explicitly profits from this as it works best on functions that are quadratic.[53] The surrogate model employed by BO assumes that the model function outputs are normally distributed. By taking the third root of a chi-squared distributed random variable it can be transformed into a more normally distributed random variable.[54] Therefore, BO minimized ($\chi^2)^{1/3}(p)$ instead of $\chi^2(p)$. Of the investigated methods only the Nelder–Mead algorithm can not take advantage of derivative information if provided.

For each reconstruction method and dataset, six consecutive optimizations were performed. From these results we calculated a mean and a standard deviation of the respective reconstruction result. The metric chosen to quantify the reconstruction performance of each method was the distance $d(p)$ of the reconstructed parameter to the LSQE point $p_{LSQE}$, given in units of the reconstructed Gaussian model parameter uncertainties $\sigma_{LSQE}$, that is,

$$d(p) = \sqrt{\sum_{i=1}^{M} \left( \frac{p_i - p_{LSQE,i}}{\sigma_{LSQE,i}} \right)^2}$$

(17)

We considered the reconstruction to be converged if parameter values with a distance $d < 0.1$ were found. That is, the parameters deviated only by 10% of the confidence interval. Note that the values for $p_{LSQE}$ and $\sigma_{LSQE}$ are determined after all optimizations are finished. $p_{LSQE}$ is determined from the smallest $\chi^2$ value for each iteration across all optimization runs and schemes in the benchmark, the value for $\sigma_{LSQE}$ is determined from the Jacobian for that parameter.

To demonstrate the positive impact of using the effective DoFs $K$ to parameterize the predictive distribution in the BTVO scheme, we performed the same benchmarks as described above, but instead using the number of data channels $K$ to parameterize the predictive distribution.

Afterward, we assessed the efficiency of surrogate model augmented MCMC sampling by applying it to the MGH17 dataset, and compared the results to those of MCMC sampling applied to the exact likelihood function. Finally, we briefly investigated the result of the surrogate augmented MCMC sampling applied to the experimental GIXRF dataset.

The reconstruction of the GIXRF problem was performed on a Dell PowerEdge R7525 Rack Server with 2x AMD EPYC 7542 32-core CPUs (yielding 128 usable threads) and 1 TB of RAM installed. The reconstructions of the two analytical problems were done on a workstation computer with an AMD Ryzen 7 3700X 8-core CPU, and 32 GB of RAM.

The Bayesian methods employed are implemented in the analysis and optimization tool kit of the commercial finite element Maxwell solver JCMsuite.[55] The research data for this manuscript will be released separately.[56]

3.1. The Experimental Dataset: Grazing Incidence X-Ray Fluorescence (GIXRF)

GIXRF[57,58] is an indirect optical measurement method that can be used to quantify samples both in terms of their geometry as well as in their material composition. To investigate a sample, it is illuminated using X-ray light. The incident radiation penetrates the sample to a certain depth that depends on the incoming angle $\theta$ (c.f. Figure 3). A large fraction of the incoming radiation is reflected. The reflected radiation interferes with the incident radiation and leads to the so-called X-ray standing wave (XSW) field. A small fraction of the incident radiation is absorbed by the sample and promptly given off again in the form of fluorescent light. Regions with constructive interference contribute more strongly to the fluorescence signal. As penetration depth and XSW depend strongly on the incidence angle $\theta$, so does also the fluorescence spectrum, which is recorded by a silicon drift detector (SDD) as shown in Figure 3. The modified 2D Sherman equation is then used to determine the geometrical (and experimental) parameters of the sample using a set of experimentally obtained fluorescence intensities, a parameterized forward model of the experimental measurement process was created. In this model, a simulation of the electromagnetic fields of the XSW is performed. The modified 2D Sherman equation is then used to determine the numerical fluorescence intensity for each angle of incidence found in the experimental dataset.[58]

The experimental dataset obtained for X-ray light with an energy of 520 eV contains 208 discrete datapoints shown in Figure 4. Note that the measurement uncertainties are approximately two orders of magnitude smaller than the measurement signal. The GIXRF reconstruction problem contains ten free parameters. Seven parameters describe the shape of the sample
The parameterized forward model was created using the finite-element discretization and a simulated electric field intensity condition in horizontal direction and perfectly absorbing boundary conditions in vertical direction. The geometric parameters are shown in white, and the remaining schemes did not manage to utilize the provided optimization budget to reconstruct parameters within 10 standard deviations of the LSQE. Listed are the height $h$, the critical dimension $cd$, the side wall angle $swa$, the thickness of the oxide layer on top of the structure $t_{\text{oxide}}$, the thickness of the oxide layer in the groove $t_{\text{groove}}$, the thickness of the substrate $t_{\text{sub}}$, the corner rounding radius $R$, a scaling parameter for the measured fluorescence $\sigma$, as well as offset values for the angle of incidence $\theta$ and the azimuth angle $\phi$, $\Delta \theta$, and $\Delta \phi$, respectively.

### Table 1.
The fitting parameters, the corresponding ranges for the GIXRF model, as well as the results of the reconstruction due to the BTVO method, that is, $p_{\text{LSQE}} \pm \varepsilon_{\text{LSQE}}$. The reconstruction results were rounded to the first significant digit of the reconstructed uncertainties.

| Parameter | Range | Reconstruction results |
|-----------|-------|------------------------|
| $h$ $[\text{nm}]$ | $[85, 100]$ | $89.5 \pm 0.4$ |
| $cd$ $[\text{nm}]$ | $[35, 55]$ | $46.3 \pm 0.2$ |
| $swa$ $[\degree]$ | $[75, 90]$ | $83.7 \pm 0.1$ |
| $t_{\text{oxide}}$ $[\text{nm}]$ | $[1, 6]$ | $2.21 \pm 0.03$ |
| $t_{\text{groove}}$ $[\text{nm}]$ | $[0.1, 10]$ | $1.0 \pm 0.3$ |
| $t_{\text{sub}}$ $[\text{nm}]$ | $[0.1, 10]$ | $6.9 \pm 0.9$ |
| $R$ $[\text{nm}]$ | $[3, 10]$ | $7.0 \pm 0.8$ |
| $\sigma_r$ $[\text{fT}]$ | $[0.5, 1.5]$ | $0.727 \pm 0.005$ |
| $\Delta \theta$ $[\degree]$ | $[-0.15, 0.15]$ | $-0.101 \pm 0.003$ |
| $\Delta \phi$ $[\degree]$ | $[-0.075, 0.075]$ | $0.006 \pm 0.009$ |

#### 3.1.1. Reconstruction Results

Figure 6 shows the performance of the employed optimization methods at reconstructing the geometrical and experimental model parameters for the GIXRF dataset. The benchmark’s mean is depicted as a solid line, while its standard deviation is shown as a shaded band around the mean. Figure 6a shows the progress of the reconstruction in terms of calls to the model function, while Figure 6b shows it in terms of the actual wall time. This differentiation highlights the slight computational overhead of the Bayesian methods, where we are focusing on comparing the BTVO and LM in particular. BTVO required $\approx 40$ iterations to cross the (arbitrarily chosen) $d = 1$ threshold, while LM required $\approx 88$ iterations, an increase of 120%. If we consider the wall time, BTVO required $\approx 32$ min to cross the $d = 1$ threshold and LM required $\approx 68$ min, which is a slightly smaller increase of 110%. For model functions with very short computation times, such as models described by analytic functions, BTVO may not necessarily be advantageous. On average, a single forward simulation costs $\approx 30$ s, while the generation of a new sample candidate with the BTVO method costs on average between 2 and 3 s. Dropping the approximation of using the same covariance kernel matrix for all GPs would lead to an increase in the sample generation of over 6 min.

Of the employed reconstruction algorithms, only the proposed BTVO and LM managed to reconstruct the parameters to an average of less than one standard deviation of the LSQE. The BTVO reached an average distance of $d \approx 0.13$, while LM reached an average distance of $d \approx 0.58$. The remaining schemes did not manage to account for uncertainties in the angles of incidence $\theta$ and $\phi$ (c.f. Figure 3). A listing of the employed parameter optimization intervals is given in Table 1. The densities of Si3N4 and SiO2, $\rho_{\text{Si3N4}} = 2.836 \text{ g cm}^{-3}$ and $\rho_{\text{SiO2}} = 2.093 \text{ g cm}^{-3}$, were obtained by means of a separate X-ray reflectometry experiment.\[58\] The parameterized forward model was created using the finite-element Maxwell solver JCMsuite.

(c.f. Figure 5), one parameter scales the calculated intensities to the experimental intensities, and two free parameters have been introduced to account for uncertainties in the angles of incidence $\theta$ and $\phi$ (c.f. Figure 3). A listing of the employed parameter optimization intervals is given in Table 1. The densities of Si3N4 and SiO2, $\rho_{\text{Si3N4}} = 2.836 \text{ g cm}^{-3}$ and $\rho_{\text{SiO2}} = 2.093 \text{ g cm}^{-3}$, were obtained by means of a separate X-ray reflectometry experiment.\[58\] The parameterized forward model was created using the finite-element Maxwell solver JCMsuite.
reconstructed uncertainties. Explanations regarding to the meaning of the parameters can be found in the table caption of Table 1, as well as in Figures 3 and 5.

The X-ray light energy of 520 eV is sufficient to obtain a fluorescence signal from the nitrogen within the core of the sample. Nevertheless, also the different oxide layer thicknesses $t_{\text{oxide}}$, $t_{\text{groove}}$, and $t_{\text{substrate}}$ could be reconstructed with small uncertainty, which is in agreement with the observation made by Soltwisch et al. [21]

### 3.2. Analytical Datasets: MGH17 and Gauss3

Additionally, two analytical datasets were investigated. Here, the MGH17 [47] and Gauss3 [48] datasets were chosen. Both were obtained from the NIST Standard Reference Database [20]. The target values within the datasets were created using the respective model function, where a normally distributed error was added to each data point. Since the model functions are analytical in nature, derivatives are easily calculated. The model function used to fit the 33 datapoints of the MGH17 dataset is

$$f(x, \beta) = \beta_1 + \beta_2e^{-x\beta_1} + \beta_3e^{-x\beta_2}$$

and contains five free parameters (c.f. Table 2). The Gauss3 dataset consists of 250 discrete datapoints and is fit using the eight parameter model function (c.f. Table 3)

$$f(x, \beta) = \beta_1e^{-x^2\beta_1} + \beta_2e^{-(x-\beta_1)^2/\beta_2} + \beta_3e^{-(x-\beta_2)^2/\beta_3}$$

#### 3.2.1. MGH17 Reconstruction Results

The results for the optimization benchmarks of the MGH17 dataset are shown in Figure 7. Figure 7a shows the results without the use of derivative information, while Figure 7b shows the results which were obtained with accurate derivative information taken into consideration. For better visibility of the convergence behavior of BTVO and LM, we have truncated the number of iterations in Figure 7b to 80. The other methods did not converge to $d < 0.1$ within an optimization budget of 350 iterations.

Without derivative information, only BTVO and LM were able to reconstruct the certified results (c.f. Table 2) within its standard deviations, as well as the uncertainty intervals. BTVO reached a value of $d < 0.1$ after $\approx 54$ iterations, while LM required 300 iterations. None of the remaining methods were able to get to within ten standard deviations of $p_{\text{LSQE}}$ before the optimization budget was exhausted.

When derivative information was taken into consideration the results for BTVO, LM, and L-BFGS-B improved. Both BTVO and LM were able to reach $d < 0.1$ (40 iterations for BTVO, 61 iterations for LM), and were ultimately also able to reconstruct the certified result within the provided optimization budget. We note that for this particular example BTVO without

### Table 2. List of parameters with certified fit results as taken from ref. [47], as well as the employed parameter bounds for the MGH17 dataset. The certified results are rounded to the first significant digit of the given uncertainty.

| Parameter | Range     | Certified value [47] |
|-----------|-----------|----------------------|
| $\beta_1$ | [0, 10]   | 0.374 ± 0.002        |
| $\beta_2$ | [0, 1.4]  | 1.9 ± 0.2            |
| $\beta_3$ | [−4, 0]   | 1.5 ± 0.2            |
| $\beta_4$ | [0.005, 0] | 0.0129 ± 0.0004      |
| $\beta_5$ | [0.005, 0] | 0.0221 ± 0.0009      |
Figure 7. The progress of the parameter reconstruction of the MGH17 dataset a) without and b) with the use of derivative information. Note that (b) shows a truncated view, limited to 80 iterations. The plots show the distance $d(p)$ to the LSQE point $p_{LSQE}$ according to Equation 17. Shown are the means (solid and dashed lines) and standard deviation (shaded bands) from six consecutive reconstruction runs. The x-axis shows the number of evaluations of the model function. Without the use of derivative information only BTVO and LM manage to reconstruct parameters that are within a distance $d < 0.1$. Taking accurate derivative information into consideration improves both the BTVO and especially the LM results. However, BTVO still outperforms LM.

Table 3. List of parameters with certified fit results as taken from ref. [48], as well as the employed parameter bounds for the Gauss3 dataset. The certified results are rounded to the first significant digit of the given uncertainty.

| Parameter | Range       | Certified value $^{(48)}$ |
|-----------|-------------|----------------------------|
| $\beta_1$ | $[90, 110]$ | $98.9 \pm 0.5$             |
| $\beta_2$ | $[0.005, 0.05]$ | $0.0109 \pm 0.0001$       |
| $\beta_3$ | $[90, 110]$ | $100.7 \pm 0.8$           |
| $\beta_4$ | $[100, 120]$ | $111.6 \pm 0.4$           |
| $\beta_5$ | $[15, 30]$  | $23.3 \pm 0.4$            |
| $\beta_6$ | $[70, 80]$  | $74 \pm 1$                |
| $\beta_7$ | $[140, 150]$ | $147.8 \pm 0.4$           |
| $\beta_8$ | $[17, 22]$  | $19.7 \pm 0.4$            |

3.2.2. Gauss3 Reconstruction Results

The results for the optimization benchmarks of the Gauss3 dataset are shown in Figure 8, where Figure 8a depicts the results without the use of derivative information, while Figure 8b shows the results that were obtained with accurate derivative information taken into consideration.

Without derivative information, BTVO again performed best and was able to reproduce the values of the certified result (c.f. Table 3) and its uncertainty intervals. In each of the six runs it was able to reconstruct parameters within less than 10% of the standard deviation of the LSQE and reached $d < 0.1$ in $\approx 38$ iterations. LM was also able to achieve $d < 0.1$, and did so after 63 iterations. The remaining three optimization schemes—BO, L-BFGS-B, and NM—were not able to get to within ten standard deviations of $p_{LSQE}$.

Taking derivative information into consideration improved the reconstruction performance for all optimization schemes. L-BFGS-B was now also able to reconstruct the model parameters to within ten standard deviations. Very drastic improvements were seen for both the proposed BTVO and LM schemes. BTVO now only required $\approx 13$ iterations to reconstruct model parameters with $d < 0.1$, approximately a third of the iterations required without derivative information. The clearest benefactor of accurate derivative information however was LM, which was now able to reconstruct the model parameters to within 10% in only seven iterations. The performance of BO increased slightly, although not to the extent that we expected when providing accurate derivative information. We attribute this again to the behavior described in Section 3.2.1. However, as the model parameters of the Gauss3 problem are correlated not as strongly as those in the MGH17...
The progress of the parameter reconstruction of the Gauss3 dataset a) without and b) with the use of derivative information. The plots show the distance $d(p)$ to the LSQE point $p_{\text{LSQE}}$ according to Equation 17. Shown are the means (solid and dashed lines) and standard deviation (shaded bands) from six consecutive reconstruction runs. The x-axis shows the number of evaluations of the model function. Both with and without derivative information, only BTVO and LM manage to reconstruct the parameters to within $d < 0.1$ in the provided optimization budget. Without derivative information BTVO performs better than LM, and with derivative information LM performs better than BTVO.

3.3. BTVO without Effective Degrees of Freedom

As detailed in Section 2.2.3, keeping the DoFs fixed to the number of data channels $K$ can lead to a very localized optimization behavior for the BTVO scheme. The associated selection of samples that are too close to previous sample positions leads to a termination of the optimization after a certain number of iterations. We have observed that model functions with a smaller number of data channels took longer to reach this point. The number of iterations taken by each model function averages from six different runs. Here, the GIXRF reconstruction terminated after an average of 14 iterations. The MGH17 reconstructions terminated after an average of 73 iterations without, and 31 iterations with the use of derivatives. The Gauss3 reconstructions terminated after an average of 14 iterations without, and only 4 iterations with the use of derivatives. During this small number of iterations, the BTVO scheme was only able to reconstruct parameters with $d \lesssim 1$ for the MGH17 problem and otherwise failed to reconstruct the parameters ($d > 1$).

3.3.1. Impact of the Effective Degrees of Freedom on the Approximated Distribution

By employing the effective DoF parameterization for the approximated probability distribution, the parameter reconstruction of the GIXRF problem was successful in each of the six instances in the benchmark Section 3.1.1. When fixing the effective DoFs to the full number of data channels $K$ we have observed that the optimizations in the benchmark terminated on average after 14 iterations, because the employed infill criterion was not able to generate new valid samples.

We consider this particular iteration for an analysis of the impact of the effective DoF parameterization. For a successful optimization we consider the known subsequent 15th iteration, that is its parameter $p_{15}$ and the associated observed model function values $f(p_{15})$. As an example, we deliberately set the surrogate uncertainty to be $\sigma = 10 \cdot \eta$ (ten times the measurement uncertainty) and calculate the approximated probability distribution using an effective DoF parameterization and using a fixed $K$ DoF parameterization, using Equations 12 and 13 (c.f. Figure 9).

We observe that both parameter distributions largely differ. While the mode of the effective DoF probability distribution is close to the minimal observed $\chi^2$ value, the probability density of
the fixed $K$ DoF distribution is almost zero for $\chi^2 \approx \chi^2_{\text{min}}$. Hence, the LCB infill criterion in the fixed $K$ case does not choose the sample $p_{\text{LS}}$ but rather parameters so close to the observed minimum of $\chi^2$ that the minimization is stopped. This exemplifies, that the fixed $K$ parameterization hinders the efficient exploration of the parameter space because it is too pessimistic about finding small $\chi^2$ values.

4. Application of Markov Chain Monte Carlo Sampling

To demonstrate the capabilities of the surrogate model augmented MCMC sampling, we applied it to the analytic MGH17 dataset. The goal was to sample the predicted likelihood function $\hat{L}(p)$ to extract the parameter uncertainties in terms of the 16%, 50% (or median), and 84% percentiles of the parameter distributions, and to determine the correlations between the parameters of the model. Because the MGH17 dataset is based on an analytic function, we also applied MCMC to the exact likelihood function $L(p)$ and compared the obtained results. To sample the respective likelihood functions we employed the Python package emcee.[60]

The surrogate model was trained in two stages. In the parameter reconstruction stage the LSQE was found within 63 iterations. Derivative information was not used during the reconstruction. In the refinement stage, the uncertainty of the surrogate model in the region of interest around the LSQE is reduced. At each refinement step, $S = 10 \times (N + 1)$ random samples were drawn and the one with the largest mean uncertainty was used to evaluate the model function and retrain the surrogate model. The refinement was stopped after the maximum mean uncertainty was below $\sigma_{\text{min}} = 1 \times 10^{-4}$ times that of the global uncertainty in the surrogate model for five consecutive iterations, which was reached after an additional 41 iterations. The actual model function was thus evaluated 104 times.

**Figure 10** shows an excerpt of the comparison of MCMC sampling of the exact likelihood function $L(a, b, p)$ and the predicted likelihood function $\hat{L}(a, b, p)$ around the LSQE $p_{\text{LS}}$. In both instances, 32 MCMC walkers were used to draw 50,000 samples from the respective likelihood functions. The complete comparison for all parameters can be found in Figure S1, Supporting Information.

The histograms at the top show the 16%, 50% (or median), and 84% percentiles of the parameter distributions. Both the exact, as well as the surrogate aided MCMC samplings produced very similar histograms, both in terms of percentile values as well as in terms of the overall shape. The recovered values matched the certified results provided by the NIST within the reconstructed standard deviations, see also Table 2.

The 2D scatter plots between the histograms highlight correlations between the model parameters. The displayed levels are—from inside to outside—the $0\sigma$, $1\sigma$, $1.5\sigma$, and $2\sigma$ regions. Again, both methods showed good agreement.

Finally, the method was used to assess the magnitude of the measurement uncertainties. To this end the basic error model

$$\hat{\eta}_i = 2^\eta_i \cdot \eta_i$$

was fit to the data. The actual measurement uncertainties can be scaled up and down by fitting the value of the exponent $\eta_i$. Both MCMC sampling of the exact model and the surrogate model yield a value of $\eta_i$ close to zero with an uncertainty of about $\pm0.2$, c.f. Figure S1, Supporting Information.

This demonstrates that qualitatively and quantitatively accurate results can be obtained with the surrogate model aided MCMC, which comes at a fraction of the cost of directly sampling the true likelihood function if the model function is expensive to evaluate.

4.1. Surrogate Augmented MCMC Convergence Behavior

The accuracy of the surrogate augmented MCMC sampling depends strongly on the accuracy of the underlying GP model
and therefore on the number of additional training samples for the GP evaluated after the optimization. In order to illustrate the convergence of the MCMC results with respect to the number of additional samples, we considered the following setup. We start from the MGH17 optimization results from Section 4 and consider refinements of the GP surrogate with \( N \in \{0, 5, 10, 20, 30, 40, 50, 75, 100, 150\} \) additional training samples. Using the refined surrogate we then drew 500 000 samples using 32 MCMC walkers. From these samples we determined the 16%, 50%, and 84% percentiles for each parameter \( \beta_i \). To create reference values we employed \texttt{emcee} to draw 500 000 samples from the exact likelihood function for six consecutive runs. For each run we then determined the same percentiles as described above, the reference percentiles were then obtained by calculating the mean values across the six runs. The number of samples drawn was increased to improve the stability of the determined percentiles. The relative deviation from the exact likelihood result was determined for each percentile and parameter \( \beta_i \) and then averaged for each \( N \) across these parameters and percentiles. The results are shown in Figure 11. To better highlight the trend, a least-squares fit trend line is shown as well. A theoretical threshold of \( \bar{\epsilon}_\text{rel. MCMC perc.} \approx 8 \times 10^{-3} \) was determined by determining the average relative standard deviation from the six reference value runs.

We observe that for an increasing number of refinement samples \( N \) the relative deviation from the exact results decreases from initially 5% to about 1%. This shows that accurate uncertainty estimates can be obtained with a relatively small number of forward model evaluations.

4.2. MCMC Sampling of the Approximate Likelihood Function of the Experimental GIXRF Dataset

Using the surrogate augmented MCMC method we have sampled the approximate likelihood function of the GIXRF dataset. An excerpt of the results is shown in Figure 12, where we have limited the displayed parameters to the critical dimension \( c_d \), the scaling parameter \( s_N \), and the offset of the incidence angle \( \theta \). For these parameters we observe correlations at the reconstructed parameter value \( p_{\text{LSQE}} \). This is due to the fact that the right-most maximum in the measurement data, c.f. Figure 4 at \( \approx 89^\circ \), is strongly correlated to the critical dimension \( c_d \). A change in the offset angle \( \Delta \theta \) moves the peak, which leads to a correlation between \( c_d \) and \( \Delta \theta \). The correlation between \( s_N \) and \( c_d \) can be explained if we consider, that a larger critical dimension means that the grating contains more material, which increases the fluorescence signal. \( s_N \) is the parameter that scales this signal to match the experimentally observed one. In order to still match the experimental measurement when increasing the critical dimension, the fluorescence signal has to be scaled down. The results shown in the histograms confirm the results obtained from the reconstruction using the BTVO method, c.f. Table 1.

The remaining model parameters showed only minor correlations.

5. Conclusion

We have expanded a recently introduced BTVO scheme to be compatible with a much larger number of data channels \( K \). We have done this by introducing a new parameterization, which adjusts the actual DoFs to be an effective number of DoFs, as well as by sharing the covariance function for the individual GPs among each other, such that only one matrix decomposition is necessary at each iteration of the optimization instead of one for each data channel.

We have applied the proposed BTVO to various parameter reconstruction tasks and have shown that the scheme regularly outperforms the established Levenberg–Marquardt algorithm at reconstructing model parameters during a least-squares fit, the single exception being when derivative information has been directly exploited in one particular example.

We have further shown that the method can be extended to enable Markov chain Monte Carlo sampling of expensive model functions with largely reduced computational costs. This has been achieved by exploiting the multi-output surrogate model
which was trained during the parameter reconstruction. By using an analytic model function we have demonstrated that the surrogate model approach is capable of generating results that are in good agreement with the exact results, while using only a fraction of the evaluations of the model function. In doing this, one is able to obtain model parameter uncertainties in terms of percentiles of the model parameter distributions, to determine non-linear correlations between parameters, and in principle also to fit a non-trivial error model to the data.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The authors acknowledge discussions with Victor Soltwisch, Martin Hammerschmidt, and Lin Zschiedrich, and also acknowledge Philipp Hönice for assistance in recording the experimental data set. This project was funded by the German Federal Ministry of Education and Research (BMBF; project number 05M20ZAA, siMLopt; project number 01IS20080A, SI4M4D; Forschungscampus MODAL, project number 05M20ZBM), by the German Federal Ministry for Economic Affairs and Energy (BMWi; project number 50WM2067, Optimal-QT), and by the European Union’s Horizon 2020 research and innovation programme (EU H2020, grant number 101007319, AI-TWILIGHT). This project has received funding from the EMPIR programme co-financed by the Participating States and from the European Union’s Horizon 2020 research and innovation programme (project 20INN04 “ATMOC”; project 20FUUN02 “POLIGHT”).

Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.6359593 https://doi.org/10.5281/zenodo.6359594, reference number 6359593.

Keywords

Bayesian target-vector optimization, least-squares, metrology, parameter reconstruction, uncertainty quantification

Received: February 22, 2022
Revised: May 23, 2022
Published online: August 24, 2022

[1] K. Levenberg, Q. Appl. Math. 1944, 2, 164.
[2] D. W. Marquardt, J. Soc. Ind. Appl. Math. 1963, 11, 431.
[3] P. Deuflhard, Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms, Vol. 35, Springer Science & Business Media, Berlin, Heidelberg 2005.
[4] E. Vogel, Nat. Nanotechnol. 2007, 2, 25.
[5] N. G. Orji, M. Badaroglu, B. M. Barnes, C. Beita, B. D. Bunday, U. Celano, R. J. Kline, M. Neisser, Y. Obeng, A. Vladar, Nat. Electron. 2018, 1, 532.
[6] C. J. Raymond, M. E. Littau, A. Chuprin, S. Ward, Proc. SPIE 2004, 5375, 564.
[7] J. Močkus, in Optimization Techniques IFIP Technical Conference, Springer, Berlin, Heidelberg 1975, pp. 400–404.
[8] J. Močkus, Bayesian Approach to Global Optimization: Theory and Applications, Vol. 37, Springer Science & Business Media, Berlin, Heidelberg 2012.
[9] C. K. Williams, C. E. Rasmussen, Gaussian Processes for Machine Learning, Vol. 2, MIT Press, Cambridge, MA 2006.
[10] D. R. Jones, M. Schonlau, W. J. Welch, J. Global Optim. 1998, 13, 455.
[11] K. Matsui, S. Kusakawa, K. Ando, K. Kutsukake, T. Ujihara, I. Takeuchi, arXiv:1103.0371, 2019.
[12] A. K. Uhrenholt, B. S. Jensen, Proc. Mach. Learn. Res 2019, 89, 2661.
[13] C. Huang, Y. Ren, E. K. McGuinness, M. D. Losego, R. P. Lively, V. R. Joseph, Optim. Eng. 2021, 22, 2553.
[14] B. M. Barnes, M.-A. Henn, Proc. SPIE 2020, 11325, 113251E.
[15] M. A. Alvarez, L. Rosasco, N. D. Lawrence, arXiv:1106.6251, 2011.
[16] H. Liu, J. Cai, Y.-S. Ong, Knowl. Based Syst. 2018, 144, 102.
[17] M. Hammerschmidt, M. Weiser, X. G. Santiago, L. Zschiedrich, B. Bodermann, S. Burger, Proc. SPIE 2017, 10330, 1033004.
[18] V. Szewdowski-Rammert, J. Baumann, C. Schlesiger, U. Waldschläger, A. Gross, B. Kanngießer, I. Mantouvalou, J. Anal. At. Spectrom. 2019, 34, 922.
[19] N. Kumar, P. Petrik, G. K. Ramanandan, O. El Gawhary, S. Roy, S. F. Pereira, W. M. Coene, H. P. Urbach, Opt. Express 2014, 22, 24678.
[20] NIST, Standard Reference Database, https://www.itl.nist.gov/div898/strd/index.html (accessed: February 2022).
[21] V. Soltwisch, P. Hönice, Y. Kayser, J. Elbracht, J. Probst, F. Scholze, B. Beckhoff, Nanoscale 2018, 10, 6177.
[22] M. Shokooh-Saremi, R. Magnusson, Opt. Lett. 2007, 32, 894.
[23] S. M. Mirjalili, K. Abedi, S. Mirjalili, Optik 2013, 124, 5989.
[24] E. Bor, M. Turducv, H. Kurt, Sci. Rep. 2016, 6, 30871.
[25] M. G. Saber, A. Ahmed, R. H. Sagor, Silicon 2017, 9, 723.
[26] A. F. Herrero, M. Pflüger, J. PulS, F. Scholze, V. Soltwisch, Opt. Express 2021, 29, 35580.
[27] M. H. Kutner, C. J. Nachtsheim, J. Neter, W. Li, Applied Linear Statistical Models, 5th ed., McGraw-Hill Irwin, New York 2005.
[28] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, Numerical Recipes: The Art of Scientific Computing, 3rd ed., Cambridge University Press, Cambridge 2007.
[29] T. Strutz, Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and Beyond, Vieweg and Teubner, Wiesbaden 2011.
[30] M. Pflüger, R. J. Kline, A. F. Herrero, M. Hammerschmidt, V. Soltwisch, M. Krumrey, J. Micro Nanolithogr. MEMS, MOEMS 2020, 19, 014001.
[31] C. Andrieu, N. De Freitas, A. Doucet, M. I. Jordan, Mach. Learn. 2003, 50, 5.
[32] P.-I. Schneider, X. Garcia Santiago, V. Soltwisch, M. Hammerschmidt, S. Burger, C. Rockstuhl, ACS Photonics 2019, 6, 2726.
[33] R. Garnett, Bayesian Optimization, Cambridge University Press, Cambridge 2022.
[34] E. Brochu, V. M. Cora, N. De Freitas, arXiv:1012.2599, 2010.
[35] X. Garcia-Santiago, P.-I. Schneider, C. Rockstuhl, S. Burger, J. Phys. Conf. Ser. 2018, 963, 012003.
[36] X. Garcia-Santiago, S. Burger, C. Rockstuhl, P.-I. Schneider, J. Light. Technol. 2021, 39, 167.
[37] A. M. Mathai, S. B. Provost, Quadratic forms in Random Variables: Theory and Applications, Marcel Dekker, New York 1992.
[38] A. A. Mohsenipour, Ph.D. Thesis, University of Western Ontario, London, ON 2012.
[39] M. Plock, S. Burger, P.-I. Schneider, *Proc. SPIE* 2021, 11783, 117830J.

[40] S. Abdel-Aty, *Biometrika* 1954, 41, 538.

[41] C. Sammut, G. I. Webb, *Encyclopedia of Machine Learning*, Springer Science & Business Media, Berlin/Heidelberg, Germany 2011.

[42] J. Friedman, T. Hastie, R. Tibshirani, T. Hastie, *The Elements of Statistical Learning*, Vol. 1, Springer Series in Statistics, Springer, New York, NY 2001.

[43] M.-A. Henn, H. Gross, F. Scholze, M. Wurm, C. Elster, M. Bär, *Opt. Express* 2012, 20, 12771.

[44] N. Farchmin, M. Hammerschmidt, P.-I. Schneider, M. Wurm, B. Bodermann, M. Bär, S. Heidenreich, *J. Micro Nanolithogr. MEMS MOEMS* 2020, 19, 024001.

[45] A. Andrle, N. Farchmin, P. Hagemann, S. Heidenreich, V. Soltwisch, G. Steidl, in *Scale Space and Variational Methods in Computer Vision* (Eds: A. Elmoataz, J. Fadili, Y. Quéau, J. Rabin, L. Simon), Springer International Publishing, Cham 2021, pp. 528–539.

[46] R. B. Gramacy, *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*, Chapman and Hall/CRC, London 2020.

[47] NIST Standard Reference Database, MGH17 dataset, https://doi.org/10.18434/T43G6C (accessed: February 2022).

[48] NIST Standard Reference Database, Gauss3 dataset, https://www.itl.nist.gov/div898/strd/nls/data/gauss3.shtml (accessed: February 2022).

[49] R. Fletcher, A Modified Marquardt Subroutine for Non-Linear Least Squares, Technical Report AERE-R-6799, Atomic Energy Research Establishment, Harwell (England) 1971.

[50] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, et al., *Nat. Methods* 2020, 17, 261.

[51] R. H. Byrd, P. Lu, J. Nocedal, C. Zhu, *SIAM J. Sci. Comp.* 1995, 16, 1190.

[52] J. A. Nelder, R. Mead, *Comput. J.* 1965, 7, 308.

[53] D. C. Liu, J. Nocedal, *Math. Program.* 1989, 45, 503.

[54] D. M. Hawkins, R. Wixley, *Am. Stat.* 1986, 40, 296.

[55] JCMwave, JCMSuite, https://jcmwave.com/jcmsuite/ (accessed: February 2022).

[56] M. Plock, K. Andrle, S. Burger, P.-I. Schneider, Zenodo, Research data and example scripts for the paper: Bayesian Target-Vector Optimization for Efficient Parameter Reconstruction, Zenodo 2022, https://doi.org/10.5281/zenodo.6359593.

[57] Y. Ménesguen, B. Boyer, H. Rotella, J. Lubeck, J. Weser, B. Beckhoff, D. Grötzsch, B. Kanngießer, A. Novikova, E. Nolot, M.-C. Lépy, *X-Ray Spectrum*. 2017, 46, 303.

[58] A. Andrle, P. Hönicke, G. Gwalt, P.-I. Schneider, Y. Kayser, F. Siewert, V. Soltwisch, *Nanomaterials* 2021, 11, 1647.

[59] P. Hönicke, A. Andrle, Y. Kayser, K. V. Nikolaev, J. Probst, F. Scholze, V. Soltwisch, T. Weimann, B. Beckhoff, *Nanotechnology* 2020, 31, 505709.

[60] D. Foreman-Mackey, D. W. Hogg, D. Lang, J. Goodman, *Publ. Astron. Soc. Pac.* 2013, 125, 306.