Using Gaussian process regression for efficient parameter reconstruction

Philipp-Immanuel Schneider, [a] Martin Hammerschmidt, [a] Lin Zschiedrich, [a] Sven Burger [ab]

[a] JCMwave GmbH
Bolivarallee 22, D – 14 050 Berlin, Germany
[b] Zuse Institute Berlin (ZIB)
Takustraße 7, D – 14 195 Berlin, Germany

ABSTRACT

Optical scatterometry is a method to measure the size and shape of periodic micro- or nanostructures on surfaces. For this purpose the geometry parameters of the structures are obtained by reproducing experimental measurement results through numerical simulations. We compare the performance of Bayesian optimization to different local minimization algorithms for this numerical optimization problem. Bayesian optimization uses Gaussian-process regression to find promising parameter values. We examine how pre-computed simulation results can be used to train the Gaussian process and to accelerate the optimization.

Keywords: computational metrology, optical metrology, computational lithography, Bayesian optimization, machine learning, finite-element methods, nanooptics

1. INTRODUCTION

Geometry reconstruction based on scatterometric data is a challenging numerical task. The sample structures and the measurement process are typically described using many parameters. This leads to high-dimensional optimization problems of finding shape parameters that are in agreement with the experimental data.[1–3]

The forward-problem of computing the scattering behavior of the setup at a given point in the parameter space requires to rigorously solve Maxwell's equations. We use our finite-element method (FEM) implementation JCMsuite[4,5] to this aim. The optimization process can be often solved more efficiently by providing parameter derivatives. Therefore, we compute the gradient of the solution of Maxwell's equations by automatic differentiation.[1]

A large number of minimization algorithms can be used to solve the inverse problem of reconstructing the shape parameters. In the considered case the objective function has only a small number of local minima such that local minimization algorithms should be very efficient. Examples for gradient-based local optimization methods are the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm and its low-memory, bound-constrained extension L-BFGS-B [6, and references therein] as well as the truncated Newton method.[7] An example for a gradient-free method is the downhill simplex algorithm (also known as Nelder-Mead method).[8]

Recently, techniques from the field of machine-learning have been employed for the optimization of photonic nanostructures. For example, deep neural networks trained with thousands of simulation results have been employed as accurate models for mapping a geometry to an optical response and vice versa almost instantaneously.[9] In this work, we consider consider Gaussian processes as a method to learn the behavior of the objective function.[10] A popular method that employs Gaussian processes is Bayesian optimization.[11] Bayesian optimization

---

is regularly used in machine learning applications.[11–13] In the field of nano-optics it has been, e.g., employed to optimize ring resonator-based optical filters[14] and chiral scatterers.[15]

Bayesian optimization derives promising parameter values by means of Bayesian inference based on *all* previous function evaluations. This is in contrast to local optimization methods, which only use few of the previous data points to determine new parameters. This statistical inference can often reduce the number of required simulations.[16]

In the context of a parameter reconstruction it is possible to compute the system response for many parameter values in advance. Provided with scatterometry data of a specific structure, the deviation between numerical and experimental response for the pre-computed parameter values can serve as training data for the Gaussian process. We investigate to which extent this training can speed up the parameter reconstruction.

2. SCATTEROMETRIC MEASUREMENT CONFIGURATION

In order to assess the optimization methods for solving the inverse problem, we use a critical dimension metrology setup studied already in previous publications.[1,17] The scatterometric measurement was executed at Physikalisch-Technische Bundesanstalt (PTB) and a detailed investigation of the data and optical model can be found in a recently published paper.[2]

For the sake of completeness, we briefly review the experimental setup and the optical model: A silicon grating (1D periodic lines) with nominal pitch of $p_x = 50$ nm and nominal line-width of $CD = 25$ nm was used as scattering target in a goniometric setup with an inspection wavelength of $\lambda = 266$ nm. A light beam with defined polarization and angle of incidence (inclination angle θ, rotation angle ϕ) illuminates the target. Due to the grating period and the wavelength only the zeroth diffraction order is present and the intensity of the reflected light in this diffraction order is recorded for S- and for P-polarized illumination at different inclination angles θ. Two azimuthal rotations ($\phi = 0$ and $\phi = 90°$) are recorded.

A schematic of the measurement is shown in Figure 1 (left). The measured data set used in this study is plotted in Figure 2 (circles). See Ref.[1] for further explanations.
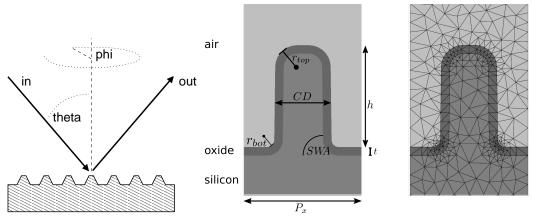

Figure 1. *Left:* Schematics of the experimental 2θ setup with incidence angle θ and azimuthal orientation ϕ. *Center:* Schematics of the model of a unit cell of the silicon line grating with free parameters line height, h, critical dimension at $h/2$, CD, oxide thickness, t, sidewall angle SWA, top and bottom corner roundings, r_{top} and r_{bot}. *Right:* Visualization of the triangular mesh for the FEM discretization.

Optical model. The aim of the optical model is to describe the measurements as a function of the microstructure's parameters $\mathbf{x} \in \mathbb{R}^n$. The shape of the silicon line is parameterized with $n = 6$ free parameters: the line height h, the line width (critical dimension) at $h/2$, CD, the oxide layer thickness, t, the sidewall angle, SWA, and the top and bottom corner roundings, r_{top} and r_{bot}. Their definitions can be found in Figure 1 (center).

In the reconstruction we allow for parameter values within large intervals describing a wide range of line shapes. To avoid non-physical self-intersections, we demand these to stay in an admissible bounded region. The admissible region can in general be defined as $A = \{\mathbf{x} \in \mathbb{R}^m \mid g_i(\mathbf{x}) \geq 0, i = 1, \ldots, r\}$ with smooth scalar valued functions g_i. For example, we demand that the corner rounding radius at the top of the line is smaller than half the width at the top. The admissible region is included into the objective function by means of a prior density as

$$\pi(\mathbf{x}) \sim \exp\left(\sum_{i=1}^{r} \log g_i(\mathbf{x})\right). \tag{1}$$

The scattering of monochromatic light off the nanoscopic line grating is described by the linear Maxwell equations in frequency domain. These lead to a single second order partial differential equation

$$\nabla \times \mu^{-1} \nabla \times \mathbf{E} - \omega^2 \varepsilon \mathbf{E} = 0, \tag{2}$$

where ε and μ are the permittivity and permeability tensors, and ω is the time-harmonic frequency. We employ the finite-element (FEM) electromagnetic field solver JCMsuite,[4,5] which has been successfully used in scatterometric investigations ranging from the optical[18] to the EUV and X-ray regimes[19,20] on 2D (e.g., line masks) and 3D (e.g., FinFETs, contact holes) scattering targets.

Inverse problem. To reconstruct likely parameters, we use the same objective function as in.[1] That is, we minimize the conditional probability $\pi(\mathbf{x}|\mathbf{y}_M)$ of the parameter vector $\mathbf{x}$ given the measurement vector $\mathbf{y}_M$, also known as posterior probability. The posterior is given by Bayes' theorem as the product of likelihood $\pi(\mathbf{y}_M|\mathbf{x})$ and prior probability $\pi(\mathbf{x})$. We assume independently normally distributed measurement errors with zero mean and diagonal covariance $\mathbf{\Gamma}_l$:

$$\pi(\mathbf{y}_M|\mathbf{x}) \sim \exp\left(-\frac{1}{2}(\mathbf{y}_M - \mathbf{y}(\mathbf{x}))^T \mathbf{\Gamma}_l^{-1}(\mathbf{y}_M - \mathbf{y}(\mathbf{x}))\right). \tag{3}$$

The posterior density

$$\pi(\mathbf{x}|\mathbf{y}_M) = \pi(\mathbf{y}_M|\mathbf{x})\pi(\mathbf{x}) \tag{4}$$

is maximized in order to find the most probable parameter values

$$\mathbf{x}_{\mathrm{MAP}} = \arg\max_{\mathbf{x} \in X} \pi(\mathbf{x}|\mathbf{y}_M). \tag{5}$$

Equivalently, $\mathbf{x}_{\mathrm{MAP}}$ can be found by minimizing the logarithm of $\pi(\mathbf{x}|\mathbf{y}_M)$,

$$F(\mathbf{x}) = \frac{1}{2}(\mathbf{y}_M - \mathbf{y}(\mathbf{x}))^T \mathbf{\Gamma}_l^{-1}(\mathbf{y}_M - \mathbf{y}(\mathbf{x})) - \sum_{i=1}^{r} \log g_i(\mathbf{x}). \tag{6}$$

As the mapping $\mathbf{y}(\mathbf{x})$ from model parameters to scatterometry measurements is nonlinear, the posterior density $\pi(\mathbf{x}|\mathbf{y}_M)$ is non-normal and can exhibit multiple local maxima. In the previous numerical study only two local maxima were found such that most of the runs of a gradient descent method converged efficiently to the global maximum $\mathbf{x}_{\mathrm{MAP}}$ located at $CD = 25.38$ nm, $h = 48.08$ nm, $SWA = 86.98°$, $t = 4.94$ nm, $r_{\mathrm{top}} = 10.37$ nm, and $r_{\mathrm{bot}} = 4.79$ nm. The local uncertainties were quantified in terms of the covariance matrix

$$\mathbf{\Gamma}_p = F''(\mathbf{x}_{\mathrm{MAP}})^{-1} \tag{7}$$

yielding the standard deviations $\sigma_{CD} = 0.395$ nm, $\sigma_h = 2.484$ nm, $\sigma_{SWA} = 0.999°$, $\sigma_t = 0.162$ nm, $\sigma_{r_{\mathrm{top}}} = 4.289$ nm and $\sigma_{r_{\mathrm{bot}}} = 3.217$ nm.[1]

In Figure 2 the experimental and simulated intensities are shown as function of inclination angle θ. The four different angular spectra refer to the different polarizations and azimuthal orientations of the illumination. We observe an almost perfect alignment of the simulated data for $\mathbf{x} = \mathbf{x}_{\mathrm{MAP}}$ and the PTB measurements.
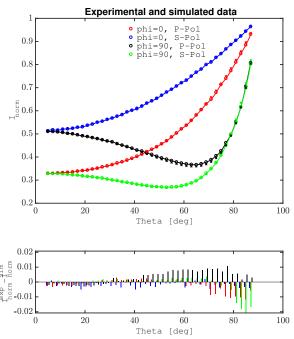
Figure 2. Experimental data (circles and connecting lines) and simulated data for the $\mathbf{x}_{\mathrm{MAP}}$ configuration (crosses). We observe a very good quantitative alignment of the data. The plot at the bottom shows the difference between measured and simulated signals. Largest deviations are observed for the S-polarization, $\phi = 90°$ and large inclination angles θ.

3. BAYESIAN OPTIMIZATION METHOD

The goal of every optimization algorithm is to identify the minimum (or maximum) of an unknown objective function f in a certain design space $\mathcal{X} \subset \mathbb{R}^d$,

$$\mathbf{x}_{min} = \arg\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}). \tag{8}$$

The basic idea of Bayesian optimization is to treat the unknown objective as a random function, i.e. a stochastic model on a continuous domain $\mathcal{X}$. Based on the previous observation of the objective the algorithm identifies parameter values where it is expected to find a smaller function value.[11,21]

Gaussian processes. Gaussian processes (GP) are frequently used as the stochastic model in Bayesian optimization. A stochastic process $(X_x)_{x \in \mathcal{X}}$ is a Gaussian process if for any N points $\mathbf{x}_1^*, \cdots \mathbf{x}_N^* \in \mathcal{X}$ the probability of the objective to be equal to $\mathbf{Y} = (y_1, \cdots, y_N)$ follows a multivariate Gaussian distribution

$$P(\mathbf{Y}^*) = \frac{1}{(2\pi)^{N/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{Y} - \mathbf{m})^T \mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{m})\right]. \tag{9}$$

with a mean vector $\mathbf{m}$ and a covariance matrix $\mathbf{\Sigma}$.[22]

The Gaussian process is defined by a covariance function (or kernel) $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a mean function $\mu : \mathcal{X} \to \mathbb{R}$. Without prior information the mean vector $\mathbf{m}$ and a covariance matrix $\mathbf{\Sigma}$ evaluate to $\mathbf{m} = [\mu(\mathbf{x}_1^*), \cdots, \mu(\mathbf{x}_N^*)]^T$ and $(\mathbf{\Sigma})_{ij} = k(\mathbf{x}_i^*, \mathbf{x}_j^*)$. After M iterations, the function values $Y = (y_1, \cdots, y_M) = (f(x_1), \cdots, f(x_M))$ are known. Using Bayesian inference, the mean vector and the covariance matrix then reads

$$\mathbf{m} = \mathbf{m}_2 + \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}(\mathbf{Y}_1^* - \mathbf{m}_1) \tag{10}$$

$$\mathbf{\Sigma} = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{21}^T \tag{11}$$

with $\mathbf{m}_1 = [\mu(\mathbf{x}_1), \cdots, \mu(\mathbf{x}_M)]^T$, $\mathbf{m}_2 = [\mu(\mathbf{x}_1^*), \cdots, \mu(\mathbf{x}_N^*)]^T$, $(\mathbf{\Sigma}_{11})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $(\mathbf{\Sigma}_{22})_{ij} = k(\mathbf{x}_i^*, \mathbf{x}_j^*)$, $(\mathbf{\Sigma}_{21})_{ij} = k(\mathbf{x}_i^*, \mathbf{x}_j)$, and $(\mathbf{\Sigma}_{12})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j^*)$.

Gaussian processes can be easily extended to incorporate not only data on the objective function but also of its derivatives.[23] For example, a gradient observation $\nabla f(\mathbf{x}_k)$ leads to additional entries in the covariance matrix formed by derivatives of the covariance function $\nabla_{\mathbf{x}_k} k(\mathbf{x}_i, \mathbf{x}_k)$ and $\nabla_{\mathbf{x}_i} \nabla_{\mathbf{x}_k} k(\mathbf{x}_i, \mathbf{x}_k)$.

From Eqs. (9–11) follows that at any position $\mathbf{x}^* \in \mathcal{X}$ the unknown function value is normally distributed, i.e. $f(\mathbf{x}^*) \sim \mathcal{N}(\bar{y}, \sigma^2)$ with

$$\bar{y}(\mathbf{x}^*) \quad = \quad \mu(\mathbf{x}^*) + \sum_{ij} k(\mathbf{x}^*, \mathbf{x}_i)(\Sigma_{11}^{-1})_{ij}[f(\mathbf{x}_j) - \mu(\mathbf{x}_j)] \tag{12}$$

$$\sigma(\mathbf{x}^*)^2 \quad = \quad k(\mathbf{x}^*, \mathbf{x}^*) - \sum_{ij} k(\mathbf{x}^*, \mathbf{x}_i)(\Sigma_{11}^{-1})_{ij} k(\mathbf{x}_j, \mathbf{x}^*). \tag{13}$$

Based on the normal distribution the next parameters values are chosen to maximize a specific acquisition function. A popular choice is the *expected improvement*, i.e. the expectation value of the improvement $\max(0, y_{\min} - f(\mathbf{x}^*))$ with respect to the currently known minimal function value $y_{\min}$

$$\alpha_{\mathrm{EI}}(\mathbf{x}^*, y_{\min}) = \mathbb{E}[\max(0, y_{\min} - f(\mathbf{x}^*))]$$
$$= \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{y_{\min} - \bar{y}(\mathbf{x}^*)}{\sqrt{2}\sigma(\mathbf{x}^*)}\right)\right](y_{\min} - \bar{y}(\mathbf{x}^*)) + \frac{\sigma(\mathbf{x}^*)}{\sqrt{2\pi}}\exp\left(-\frac{(y_{\min} - \bar{y}(\mathbf{x}^*))^2}{2\sigma(\mathbf{x}^*)^2}\right). \tag{14}$$

Other common acquisition functions are the *probability of improvement* or the *lower confidence bound*.[11]

Hyperparameter choice. Using different covariance functions $k(\mathbf{x}, \mathbf{x}')$, Gaussian processes allow to approximate a large class of random functions. A popular choice is the Matérn 5/2 covariance function

$$k(\mathbf{x}, \mathbf{x}') = s^2\left(1 + \sqrt{5}r(\mathbf{x}, \mathbf{x}') + \frac{5}{3}r(\mathbf{x}, \mathbf{x}')^2\right)\exp\left(-\sqrt{5}r(\mathbf{x}, \mathbf{x}')\right) \quad \text{with} \quad r(\mathbf{x}, \mathbf{x}')^2 = \sum_i \frac{(\mathbf{x}_i - \mathbf{x}_i')^2}{l_i^2}. \tag{15}$$

Moreover, we choose a constant mean function $\mu(\mathbf{x}) = \mu_0$. The values of the hyperparameters $\mathbf{w} = (\mu 0, s^2, l_1^2, \cdots, l_d^2)$ are essential for the performance of the optimization procedure. The idea is to maximize the likelihood $P(\mathbf{Y}) = P_{\mathbf{w}}(\mathbf{Y})$ of all known objective function values with respect to the hyperparameters. This hyper parameter optimization is computationally expensive and is only performed if the derivatives of $P_{\mathbf{w}}(\mathbf{Y})$ with respect to the length scales exceed a certain threshold.[24]

Learning from offline calculations. In a typical parameter reconstruction setup many specimens with the same type of geometry have to be probed. Therefore, we examine whether the reconstruction process can be accelerated by pre-comuting the optical model $\mathbf{y}(\mathbf{x})$ for many parameter values $\mathbf{x} \in X$. That is, provided with a measurement result $\mathbf{y}_M$ the parameters X and the corresponding posterior probabilities $P_0 = \{\pi(\mathbf{x}|\mathbf{y}_M) \mid \mathbf{x} \in X\}$ and the parameter derivatives $P_i = \{\partial\pi(\mathbf{x}|\mathbf{y}_M)/\partial\mathbf{x}_i \mid \mathbf{x} \in X\}$ for $i = 1, \cdots, 6$ are used to initialize the Gaussian process underlying Bayesian optimization. We draw X from a pseudo-random Sobol sequence and use only parameters that meet the constraints, i.e. $X \subset A$.[25, 26]

4. RESULTS

In order to assess the performance of the different optimization approaches we have conducted six independent optimization runs of 150 iterations with different initial conditions for each method. That is, the Downhill-Simplex method, L-BFGS-B, Newton Conjugate-Gradient, and Bayesian optimization were started from six different random initial points. Whenever a local minimization method converged to a local minimum, it was restarted at a different position. Moreover, six independent sets of points $(X^{(i)})_{i=1,\cdots,6}$ of 100 training samples and corresponding posterior probabilities and derivatives $(P_0^{(i)}, \cdots, P_6^{(i)})_{i=1,\cdots,6}$ were prepared. The sets were

used to initialize the Gaussian processes ("Bayesian optimization + training"). Apart from the downhill simplex methods, all methods make use of derivative information to determine the next sampling point.

Since the error function $F(\mathbf{x})$ [see Eq. (6)] varies between several orders of magnitude, we minimize its decadic logarithm $\lg[F(\mathbf{x})]$. Figure 3 compares the performance of the optimization approaches for minimizing $\lg[F(\mathbf{x})]$. On the left, the average objective value is shown as a function of the number of simulations for each of the different optimization methods. Because some optimization runs can fail by being trapped into a local minimum, the average objective value is not always meaningful. Therefore, the right plot shows also the median number of simulations needed to obtain objective values below a certain threshold. The median is less sensitive against failed runs, as long as at least four of the 6 runs are successful. Bayesian optimization needs on average significantly less simulations to find the minimum than the local optimization methods. Surprisingly, the performance of the derivative-free downhill-simplex method is similar to the gradient descent methods in minimizing $\lg[F(\mathbf{x})]$. Only close to the global minimum L-BFGS-B converges better to objective values below 1.1. Clearly, the 100 training samples lead to a significant speed-up of Bayesian optimization. Only after about 20 iterations the non-trained approach converges faster to the global minimum. We attribute this to the following behavior: Both Bayesian optimization methods converge into a region close to the global minimum. Because of small parameter derivatives, the expected improvement in this region becomes very small. Provided with many training samples the second approach now identifies other regions in the parameter space where some improvement can be expected and tends to explore these regions. If no training samples are available, the same happens at a later stage.

Figure 4 shows the maximum distance to the global minimum as a function of the number of simulations. The distance is measured in units of the measurement uncertainty for all six geometry parameters, i.e.,

$$d = \max_{i=1,\cdots,6} \frac{|(\mathbf{x} - \mathbf{x}_{\mathrm{MAP}})_i|}{\sigma_i} \tag{16}$$

Bayesian optimization with training converges after a median number of 9 iterations to a region within the measurement uncertainty ($d = 1$), i.e. where the derivatives of $F(\mathbf{x})$ become small. Without training, Bayesian optimization needs more than 25 iterations to converge to the same accuracy level. From the local minimization methods, only L-BFGS-B converges to the measurement uncertainty within 150 simulations.
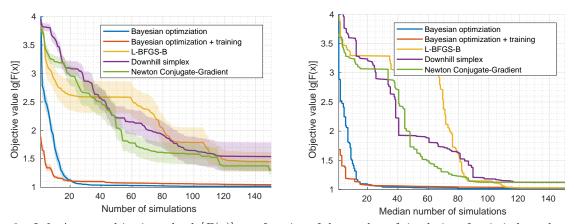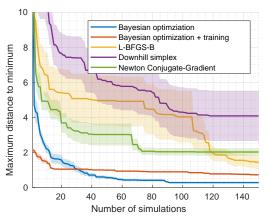

Figure 3. *Left:* Average objective value $\lg[F(\mathbf{x})]$ as a function of the number of simulations for six independent optimization runs for each method. The shaded areas indicate the standard deviation between the six optimization runs. *Right:* Median number of simulations needed to obtain an objective value smaller or equal to $\lg[F(\mathbf{x})]$ shown on the y-axis.

5. CONCLUSION

We have compared the performance of Bayesian optimization to different local minimization algorithms for a specific example of a geometry parameter reconstruction based on scatterometry data. We find that Bayesian
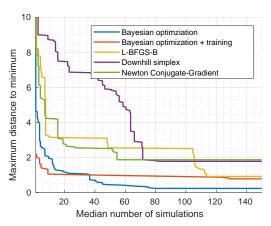
Figure 4. *Left:* Average maximum distance to global minimum as a function of the number of simulations for six independent optimization runs for each method. The shaded areas indicate the standard deviation between the six optimization runs. *Right:* Median number of simulations needed to converge to a region with a specific maximum distance to the global minimum shown on the y-axis.

optimization finds the most probable parameters with a significantly smaller number of simulation results. A training with pre-computed simulations can further speed up the reconstruction up to the point where parameter values within the measurement uncertainty are identified. Thereafter, the convergence speed decreases since the expected improvement around other training samples is larger. Other acquisition functions such as the probability of improvement could prevent the exploration of other parts of the parameter space. In order to improve the convergence speed it might be beneficial to change to a different Bayesian optimization strategy or to a gradient descent method when the expected improvement becomes small.

Acknowledgments

REFERENCES

[1] Hammerschmidt, M., Weiser, M., Santiago, X. G., Zschiedrich, L., Bodermann, B., and Burger, S., "Quantifying parameter uncertainties in optical scatterometry using Bayesian inversion," *Proc. SPIE* **10330**, 1033004 (2017).

[2] Wurm, M., Endres, J., Probst, J., Schoengen, M., Diener, A., and Bodermann, B., "Metrology of nanoscale grating structures by UV scatterometry," *Opt. Express* **25**, 2460 (2017).

[3] Soltwisch, V., Fernández Herrero, A., Pflüger, M., Haase, A., Probst, J., Laubis, C., Krumrey, M., and Scholze, F., "Reconstructing detailed line profiles of lamellar gratings from GISAXS patterns with a Maxwell solver," *J. Appl. Crystallogr.* **50**(5), 1524–1532 (2017).

[4] JCMwave GmbH, "Program package JCMsuite." http://www.jcmwave.com.

[5] Pomplun, J., Burger, S., Zschiedrich, L., and Schmidt, F., "Adaptive finite element method for simulation of optical nano structures," *Phys. Stat. Sol. (B)* **244**, 3419 (2007).

[6] Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C., "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing* **16**(5), 1190–1208 (1995).

[7] Nocedal, J. and Wright, S. J., "Large-scale unconstrained optimization," *Numerical Optimization* , 164–192 (2006).

[8] Nelder, J. A. and Mead, R., "A simplex method for function minimization," *The Computer Journal* **7**(4), 308–313 (1965).

[9] Malkiel, I., Mrejen, M., Nagler, A., Arieli, U., Wolf, L., and Suchowski, H., "Plasmonic nanostructure design and characterization via deep learning," *Light: Science & Applications* **7**(1), 60 (2018).

[10] Williams, C. K., "Prediction with Gaussian processes: From linear regression to linear prediction and beyond," in [*Learning in graphical models*], 599–621, Springer (1998).

[11] Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N., "Taking the human out of the loop: A review of Bayesian optimization," *Proceedings of the IEEE* **104**(1), 148–175 (2016).

[12] Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J., and Sculley, D., "Google vizier: A service for black-box optimization," in [*Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*], 1487–1495, ACM, New York, NY, USA (2017).

[13] "How hyperparameter tuning works on amazon sagemaker." `https://docs.aws.amazon.com/sagemaker/latest/dg/automatic-model-tuning-how-it-works.html` (2018).

[14] Rehman, S. U. and Langelaar, M., "System robust optimization of ring resonator-based optical filters," *J. Light. Technol.* **34**(15), 3653–3660 (2016).

[15] Gutsche, P., Schneider, P.-I., Burger, S., and Nieto-Vesperinas, M., "Chiral scatterers designed by Bayesian optimization," *J. Phys. Conf. Ser.* **963**(1), 012004 (2018).

[16] Schneider, P.-I., Santiago, X. G., Soltwisch, V., Hammerschmidt, M., Burger, S., and Rockstuhl, C., "Benchmarking five global optimization approaches for nano-optical shape optimization and parameter reconstruction," *arXiv preprint arXiv:1809.06674* (2018).

[17] Hammerschmidt, M., Schneider, P.-I., Santiago, X. G., Zschiedrich, L., Weiser, M., and Burger, S., "Solving inverse problems appearing in design and metrology of diffractive optical elements by using Bayesian optimization," in [*Computational Optics II*], **10694**, 1069407, International Society for Optics and Photonics (2018).

[18] Potzick, J., Dixson, R., Quintanilha, R., Stocker, M., Vladar, A., Buhr, E., Häßler-Grohne, W., Bodermann, B., Frase, C., and Bosse, H., "International photomask linewidth comparison by NIST and PTB," *Proc. SPIE* **7122**, 71222P (2008).

[19] Scholze, F., Laubis, C., Dersch, U., Pomplun, J., Burger, S., and Schmidt, F., "The influence of line edge roughness and CD uniformity on EUV scatterometry for CD characterization of EUV masks," *Proc. SPIE* **6617**, 66171A (2007).

[20] Soltwisch, V., Haase, A., Wernecke, J., Probst, J., Schoengen, M., Burger, S., Krumrey, M., and Scholze, F., "Correlated diffuse x-ray scattering from periodically nanostructured surfaces," *Phys. Rev. B* **94**, 035419 (2016).

[21] Schneider, P.-I., Santiago, X. G., Rockstuhl, C., and Burger, S., "Global optimization of complex optical structures using Bayesian optimization based on Gaussian processes," *Proc. SPIE* **10335**, 103350O (2017).

[22] Rasmussen, C. E., "Gaussian processes in machine learning," in [*Advanced lectures on machine learning*], 63–71, Springer (2004).

[23] Solak, E., Murray-Smith, R., Leithead, W. E., Leith, D. J., and Rasmussen, C. E., "Derivative observations in Gaussian process models of dynamic systems," in [*Advances in neural information processing systems*], 1057–1064 (2003).

[24] Santiago, X. G., Schneider, P.-I., Rockstuhl, C., and Burger, S., "Shape design of a reflecting surface using Bayesian optimization," *J. Phys.: Conf. Ser.* **963**, 012003 (2018).

[25] Sobol', I. M., "On the distribution of points in a cube and the approximate evaluation of integrals," *Zh. Vychisl. Mat. Mat. Fiz.* **7**(4), 784–802 (1967).

[26] Chisari, C., "Sobol: The Sobol quasirandom sequence." `http://people.sc.fsu.edu/~jburkardt/py_src/sobol/sobol.html` (2014).