Batch Constrained Bayesian Optimization for Ultrasonic Wire Bonding Feed-forward Control Design

Michael Hesse[1,3], Matthias Hunstig[2], Julia Timmermann[3] and Ansgar Trächtler[1,3]
[1]*Fraunhofer Institute for Mechatronical Systems Design, Zukunftsmeile 1, 33102 Paderborn, Germany*
[2]*Hesse GmbH, Lise-Meitner-Straße 5, 33104 Paderborn, Germany*
[3]*Heinz Nixdorf Institute, University of Paderborn, Fürstenallee 11, 33102 Paderborn, Germany*

Keywords: Bayesian Optimization, Wire Bonding, Feed-forward Control, Model-free Design.

Abstract: Ultrasonic wire bonding is a solid-state joining process used to form electrical interconnections in micro and power electronics and batteries. A high frequency oscillation causes a metallurgical bond deformation in the contact area. Due to the numerous physical influencing factors, it is very difficult to accurately capture this process in a model. Therefore, our goal is to determine a suitable feed-forward control strategy for the bonding process even without detailed model knowledge. We propose the use of batch constrained Bayesian optimization for the control design. Hence, Bayesian optimization is precisely adapted to the application of bonding: the constraint is used to check one quality feature of the process and the use of batches leads to more efficient experiments. Our approach is suitable to determine a feed-forward control for the bonding process that provides very high quality bonds without using a physical model. We also show that the quality of the Bayesian optimization based control outperforms random search as well as manual search by a user. Using a simple prior knowledge model derived from data further improves the quality of the connection. The Bayesian optimization approach offers the possibility to perform a sensitivity analysis of the control parameters, which allows to evaluate the influence of each control parameter on the bond quality. In summary, Bayesian optimization applied to the bonding process provides an excellent opportunity to develop a feed-forward control without full modeling of the underlying physical processes.

1 INTRODUCTION

Ultrasonic wire bonding (Harman, 2010) is a method of making electrical interconnections in micro and power electronics and batteries. The quality of a bonding process is specified by the so-called process capability index, which is a common performance measure in the industrial environment. Specifically for the bonding process, the process capability index depends on empirical measurements of the maximum shear force of a bond, which can be influenced by the control inputs. The goal in setting up the process is to select control inputs to maximize the bond quality while meeting certain constraints. Since the bonding process is very complex and, therefore, physically difficult to model, we apply the data-driven Bayesian optimization (BO) method from the field of machine learning to perform a feed-forward control design.

The feed-forward control design is usually realized in a model-based fashion, where the system dynamics are described by a set of physically motivated nonlinear differential equations. Considering the background of the bonding process, one of these equations should represent the time evolution of the shear force. This allows the shear force at the end of the trajectory to be determined. Based on this model, a control system can be calculated that maximizes the shear force, for example by transcription methods (Kelly, 2017).

Various publications deal with the physical modeling of the ultrasonic wire bonding process, (Mayer and Schwizer, 2002; Gogh et al., 2020; Schemmel et al., 2020) to name just a few. A common assumption is that the bond strength can essentially be described by the frictional energy induced over the process duration. These models provide a good physical explanation of how the strength of the connection increases over time. However, they lack extensive validation by measurements with various control inputs. They also have a long simulation time because the

Hesse, M., Hunstig, M., Timmermann, J. and Trächtler, A.
Batch Constrained Bayesian Optimization for Ultrasonic Wire Bonding Feed-forward Control Design.
DOI: 10.5220/0010806600003122
In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2022)*, pages 383-394
ISBN: 978-989-758-549-4; ISSN: 2184-4313

system is excited at high frequency and the connection area usually has to be discretized with a fine grid. These shortcomings make it difficult to further design the control for the bonding process. An exception is (Unger et al., 2018), where the feed-forward control is designed through multi-objective optimization for a detailed physical model. However, extensive validation is also lacking here and generalizability has not been investigated. To the best of our knowledge, there is no other practical application of model-based feed-forward control design in the literature. So far, there are no publications on the prediction of the constraints that must be satisfied in order to solve the designing problem completely model-based.

For this reason, in practice, a parameterized function is assumed for the control inputs and the dedicated parameters are manually identified through experiments on the real system, also called trial and error learning. Each time the environmental conditions of the process change, e.g. due to different materials of the contact partners, the parameters have to be re-identified. Expert knowledge that has been acquired over many years is required for this identification. Additionally, the time demand is high, because the objective function is multi modal, noisy and multiple control parameters must be set accordingly. For a human being, even a rather low-dimensional search space seems large, however, a good overview of the performed experiments is important in order to avoid redundant or irrelevant experiments. Therefore, the risk of finding only a locally optimal solution, which does not meet the quality requirements, is high. For these disadvantageous reasons, we investigate the use of BO to find the optimal control parameters for the bonding process. Our aim is to find the global maximum efficiently and robustly with an automated, objective and structured algorithm.

One of the first publications that includes the combination of BO and control theory is (Calandra et al., 2016), where BO is used to find control parameters for a walking robot. The control parameters are in this case transition times in a finite-state machine. After a relatively low number of experiments, robust walking can be realized with BO, whereas other approaches, like random search, fail to solve this task. In (Neumann-Brosig et al., 2020) BO is used to fine-tune an active disturbance rejection controller for a throttle valve which is part of combustion engines. Robustness and the duration from one desired state to another are objectives for the system. In this case remarkable results can also be achieved via the usage of BO.

The contribution of our paper is as follows: 1.) We develop an adjusted version of BO for the application

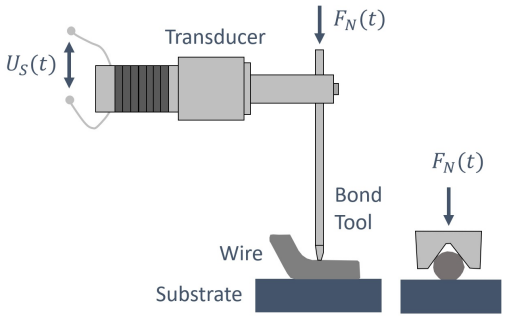

Figure 1: Components of the ultrasonic wire bonding process, side view (left, not to scale) and cross section detail (right). Control inputs are the normal force $F_N(t)$ pressing the wire to the substrate, and the alternating voltage $U_S(t)$ applied to the piezoelectric transducer, exciting in mechanical vibration.

to the bonding process which we describe in depth in Section 2. In this context, we propose adjustments to represent the process capability index and to be able to run multiple experiments (called a batch) in one iteration and to preserve the constraints. These theoretical considerations are presented in Section 3. 2.) We apply our developed method to the real bond process in Section 4 and discuss the results in Subsection 4.2. Thus, we answer the question of what kind of objective function a physical model should imply so that it can be used as prior knowledge for BO. 3.) Finally, we investigate the efficiency and robustness of our method in a simulated environment and evaluate the influence of the control inputs on the bond quality in Subsection 4.3. We conclude with a summary and outlook of future work in Section 5.

2 ULTRASONIC WIRE BONDING

Ultrasonic wire bonding is a solid-state joining process. It is a standard technology for the production of electrical interconnections in micro and power electronics for diverse applications (Harman, 2010) and also used in battery production (Hunstig et al., 2020).

Figure 1 shows the main components of an ultrasonic wire bonding process. Oscillating relative motion between wire and substrate is induced by an alternating voltage $U_S(t)$ at ultrasonic frequencies, commonly 40 to 150 kHz, applied to a tranducer containing piezoelectric elements. The transducer converts electrical excitation to mechanical vibration, which is transmitted to the process zone through a bond tool. This bond tool presses the wire to the substrate with a normal force $F_N(t)$. The two metallic partners, e.g. aluminum wire on a gold-plated substrate, are connected without melting by interdiffusion and forma-

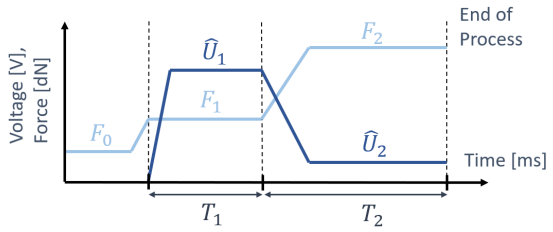

Figure 2: Assumed control function for normal force and voltage amplitude.

tion of intermetallic compounds, induced by the ultrasonic vibration. Vibration is applied for a process time from less than 10 to some 100 ms, depending on wire diameter. Usually, the transducer is held in resonance during this process by an underlying frequency controller to obtain the maximal amplitude of the tool. We assume that this frequency controller is given and appropriately tuned.

After this first bond, a wire loop to a second location is formed in a normal wire bonding process. The wire is also bonded to this destination location and afterwards, unless more than two locations shall be connected, it is severed by cutting or tearing. In this investigation, we focus on the ultrasonic bonding process of the first bond. In the experiments, we cut the wire after the first bond and do not form loops, see Figure 3. We also focus on aluminum for both the wire and the substrate in the experiments. Our feed-forward control design is governed by the normal force $F_N(t)$ and the voltage amplitude $\hat{U}_S(t)$. In Figure 2 we see the proposed parameterized control function $u(t;\theta) = [F_N(t;\theta), \hat{U}_S(t;\theta)]^T$ for both inputs. The exact shape of the control is characterized by the parameter vector $\theta = [F_0, F_1, F_2, \hat{U}_1, \hat{U}_2, T_1, T_2]^T \in \mathbb{R}^7_+$. The transition times (ramp lengths) are set to 25 % of the respective total phase time T_1/T_2 for simplicity. Physically, the bonding process consists of four phases, with a seamless overlapping transition between the last three (Geißler, 2009; Long et al., 2017): In the first *pre-deformation* phase, the wire is first pressed onto the substrate and deformed without the application of vibration. When the bond tool begins to vibrate, it causes relative motion between wire and substrate. This removes contamination such as dust particles or oxide layers and also reduces the roughness of both surfaces, it is a phase of *cleaning and activation*. After this there is a phase of large plastic *deformation* of the wire, supported by the ultrasonic softening effect (Unger et al., 2014). In the final *diffusion* phase, the contact area increases and both contact partners diffuse into another until the end of the process, resulting in a solid bond.

A good control function for creating stable bond connections supports the formation of these physical

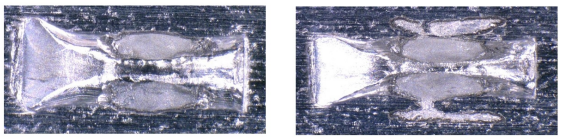

Figure 3: Microscopic images of two bonds from top view. The left connection meets all visual criteria and has a high quality, whereas the right one shows tool collisions on its sides and thus has a poor quality.

phases, but it does not have to directly reflect them in its parameters. Based on expert experience, our control function has three phases. The first phase is the pre-deformation phase, with the force F_0, no vibration and irrelevant duration. The second control phase with parameters $T_1, F_1, \hat{U}_1$ can be assumed to roughly cover the the physical processes of cleaning and activation, and some first deformation and interdiffusion. The last control phase with parameters $T_2, F_2, \hat{U}_2$ covers the rest of the deformation and most interdiffusion.

There are several criteria that define the quality of a bond. According to (DVS - German Welding Society, 2017), we particulary consider the shear strength and optical criteria, such as collisions of the tool with the substrate. The shear strength of a bond can be measured using a load cell combined with a chisel. The chisel moves over the bond at a certain small height above the substrate, meanwhile the maximum force $F_S(\theta)$ up to the breaking point of the bond is measured. This is therefore a destructive measurement method. A contact between tool and substrate can occur if the used control introduces too much energy into the system. In this case, the wire deforms too much and the edges of the bond tool collide with the substrate, possibly damaging both components. This scenario should be avoided, especially when bonding on sensitive, crack-prone substrates such as semiconductor chips. Collisions are detected with an optical microscope after the process (see Figure 3).

The process capability index C_{pK} is an important statistical measure in the industrial environment and is used in our context to quantify the quality of the bonding process. It depends on the mean and variance of the shear force, which is influenced by process and measurement noise. Thus, the shear force is a random variable with mean $\mathbb{E}[F_S(\theta)]$ and variance $\mathbb{V}[F_S(\theta)]$. The capability index is then defined by

$$C_{pK}(\theta) = \frac{\mathbb{E}[F_S(\theta)] - \mathrm{LSL}}{3 \cdot \sqrt{\mathbb{V}[F_S(\theta)]}}, \qquad (1)$$

where LSL is the lower specification limit. It determines the minimum shear force that should be achieved and is chosen depending on the application.

In order to calculate the C_{pK} value, we approximate the mean and variance empirically via

$$\mathbb{E}[F_S(\theta)] \approx \frac{1}{n_{rep}} \sum_{i=1}^{n_{rep}} F_S^{(i)}(\theta) =: \mu_{F_S}(\theta),$$
$$\mathbb{V}[F_S(\theta)] \approx \frac{1}{n_{rep}-1} \sum_{i=1}^{n_{rep}} (F_S^{(i)}(\theta) - \mu_{F_S}(\theta))^2 =: \sigma_{F_S}^2(\theta), \tag{2}$$

where we use n_{rep} separate bonds with the same underlying control, resulting in the data $F_S^{(i)}(\theta)$, with $i = 1, ..., n_{rep}$.

Tool collisions and other optical criteria are captured by the binary variable g, where 0 represents a good bond and 1 represents a bond with an optical deficit. More specifically, if at least one of the n_{rep} bonds has an optical deficit, we set g to 1.

The feed-forward control design for the ultrasonic wire bonding process can then be formulated as the following optimization problem

$$\theta_* = \arg\max_{\theta} C_{pK}(\theta), \text{ s.t. } g(\theta) = 0. \tag{3}$$

This problem is usually solved in practice by manual trial and error, as there is no automated solution yet. In the next chapter, we present our approach, which is based on Gaussian process regression and Bayesian optimization.

3 DATA-DRIVEN FEED-FORWARD CONTROL DESIGN

Before considering the application of BO (Shahriari et al., 2016), (Snoek et al., 2012), (Jones et al., 1998) to the ultrasonic wire bonding process, the methodology is explained here in detail and BO is adapted to the requirements. At the end of this section, we present the overall algorithm for the batch constrained BO. Accordingly, the section is structured as follows: Subsection 3.1 describes the construction of Gaussian process regression models. These are used to learn the unknown objective and constraint from pointwise evaluations obtained through experiments. On this basis, the batch constrained version of BO is introduced and explained in detail in Subsection 3.2.

3.1 Gaussian Process Regression

From the experiments performed, we obtain an approximation of the mean $\mu_{F_S}(\theta)$ and standard deviation $\sigma_{F_S}(\theta)$ of the shear force $F_S(\theta)$, along with the evaluation of the constraint $g(\theta)$. In the context of our approach, we treat these function values as parameter-dependent random variables, each following a particular probability distribution. The assumption is that each distribution is equal to a separate Gaussian process (Rasmussen and Williams, 2006). In the following, we present the underlying equations for the Gaussian process depending on the variable $y(\theta) \in \mathbb{R}$, which stands for one of the three functions we are looking for. For the remainder of this paper, we distinguish between the true unknown function $y(\theta)$ describing a real life process and the belief about the function that we describe with an approximation $\hat{y}(\theta)$.

A Gaussian process is completely specified by its mean $m(\theta) = \mathbb{E}[\hat{y}(\theta)]$ and its covariance function $k(\theta, \theta') = \mathbb{E}[(\hat{y}(\theta) - m(\theta))(\hat{y}(\theta') - m(\theta'))]$. Thus, the Gaussian process formally forms a distribution over the function values

$$\hat{y}(\theta) \sim \mathcal{GP}(m(\theta), k(\theta, \theta')). \tag{4}$$

With this approach there is now the possibility to appropriately choose the mean value and the covariance function for the considered process. A common assumption is that there is no prior knowledge about the process. In this case, the prior mean function is set to zero or a constant value (and thus does not depend on the parameters). However, if expert knowledge is available in advance or experiments have already been performed, a-priori knowledge about the function can be integrated into the Gaussian process framework via the mean function. We will deal with this case in Subsection 4.2. The covariance function encodes other assumptions, such as the degree of smoothness or periodicity of the unknown function. Although there are many functions and combinations of them that can be used, we figured out that the so-called 3/2-Matérn kernel (Rasmussen and Williams, 2006)

$$k(\theta, \theta'; \eta) = \sigma_f^2 (1 + \sqrt{3}d) \exp(-\sqrt{3}d) + \delta_d \sigma_n^2,$$
$$d(\theta, \theta'; \eta) = \sqrt{\sum_{i=1}^{n_\theta} \frac{(\theta_i - \theta_i')^2}{l_i^2}}, \tag{5}$$
$$\delta_d = \begin{cases} 1 & \text{if } d = 0, \\ 0 & \text{else,} \end{cases}$$

works reasonably well for our setting. The reason relies in the weak smoothness assumption implied by this type of kernel, so that even highly fluctuating functions can be reproduced. We will use this covariance function throughout this paper. The covariance function (5) depends on two configurations of the control parameters (θ, θ') and on the hyperparameters $\eta = [\sigma_f^2, l_1, ..., l_{n_\theta}, \sigma_n^2]^T \in \mathbb{R}_+^{n_\theta+2}$, which

contains the variance of the function σ_f^2, a scaling factor l_i for each dimension of the control parameters, and the variance of the noise σ_n^2. Given a set of n_d observations $D_\theta = [\theta_1, ..., \theta_{n_d}]^T \in \mathbb{R}^{n_d \times n_\theta}$, $D_y = [y_1, ..., y_{n_d}]^T \in \mathbb{R}^{n_d}$, i.e., the experimental data, these hyper-parameters can be estimated by maximizing the logarithm of the likelihood

$$\begin{aligned}\eta_* &= \arg\max_\eta \log(p(D_y)) \\ &= \arg\max_\eta -\tfrac{1}{2}(D_y - m_D)^T K(\eta)^{-1}(D_y - m_D) \\ &\quad - \tfrac{1}{2}\log(|K(\eta)|),\end{aligned} \tag{6}$$

where $m_D = [m(\theta_1), ..., m(\theta_{n_d})]^T$ and the matrix $K(\eta) \in \mathbb{R}^{n_d \times n_d}$ is the symmetric and positive definite Gram matrix with entries $K_{i,j} = k(\theta_i, \theta_j; \eta), i, j = 1, ..., n_d$. Since the logarithm is a strictly monotonously increasing function, it does not change the location of the optimum, but the numerically unstable multiplication of two possibly very small numbers becomes more stable.

The a-posteriori distribution is calculated by Bayes' theorem which is by definition a normal distribution that depends on arbitrary control parameters. The following equations describe how this distribution can be evaluated to get an estimation for mean and variance

$$\begin{aligned}p(\hat{y}(\theta) \mid D_y) &= \mathcal{N}(\mu(\theta), \sigma^2(\theta)), \\ \mu(\theta) &= m(\theta) + k_D^T(\theta) K^{-1}(D_y - m_D), \\ \sigma^2(\theta) &= k(\theta, \theta) - k_D^T(\theta) K^{-1} k_D(\theta),\end{aligned} \tag{7}$$

where we use $k_D(\theta) = [k(\theta, \theta_1; \eta_*), ..., k(\theta, \theta_{n_d}; \eta_*)]^T$.

At this point, we return to the previous functions. Thus, we have three Gaussian processes for the mean and standard deviation of the shear force and for the constraint

$$\begin{aligned}\hat{\mu}_{F_S}(\theta) &\sim \mathcal{GP}(m_\mu(\theta), k_\mu(\theta, \theta')), \\ \hat{\sigma}_{F_S}(\theta) &\sim \mathcal{GP}(m_\sigma(\theta), k_\sigma(\theta, \theta')), \\ \hat{g}(\theta) &\sim \mathcal{GP}(m_g(\theta), k_g(\theta, \theta')), \\ p(\hat{\mu}_{F_S}(\theta) \mid D_\mu) &= \mathcal{N}(\mu_\mu(\theta), \sigma_\mu^2(\theta)), \\ p(\hat{\sigma}_{F_S}(\theta) \mid D_\sigma) &= \mathcal{N}(\mu_\sigma(\theta), \sigma_\sigma^2(\theta)), \\ p(\hat{g}(\theta) \mid D_g) &= \mathcal{N}(\mu_g(\theta), \sigma_g^2(\theta)).\end{aligned} \tag{8}$$

For our experiments in Section 4, we used the Matérn kernel from (5) for all covariance functions. Although the same kernel function was used for all covariance functions, they still differ because other hyper-parameters are determined by (6) for each covariance function. The mean functions with respect to the variance of the shear force σ_{F_S} and the constraint g are set to constant values $m_\sigma = 60, m_g = 0$. In the case of the constraint, our assumption is comparable to an optimistic initialization, since we assume that the constraint is not violated in the entire parameter space. For the mean function with respect to μ_{F_S}, we consider two cases in this paper. In the first case, we set $m_\mu = \text{LSL} = 2500$ also constant and assume that there is no special prior knowledge. In the second case, we use a quadratic function $m_\mu(\theta) = \theta^T A\theta + b^T\theta + c$ instead of a constant one, where the quantities A, b, c are fitted to data via least-squares regression (details in Subsection 4.2).

Since we want to optimize the process capability index

$$\hat{C}_{pK} = \frac{\hat{\mu}_{F_S}(\theta) - \text{LSL}}{3\hat{\sigma}_{F_S}(\theta)}, \tag{9}$$

we need to combine the associated Gaussian processes for $\hat{\mu}_{F_S}$ and $\hat{\sigma}_{F_S}$. Because $\hat{C}_{pK}$ depends nonlinearly on these quantities, the related probability distribution $p(\hat{C}_{pK})$ is not Gaussian anymore. Nevertheless, the exact distribution can be calculated analytically (Díaz-Francés and Rubio, 2013). In general it is heavily tailed and has no moments. The shape can be uni modal, bi-modal, symmetric or asymmetric. However, the authors of (Díaz-Francés and Rubio, 2013) suggest a normal approximation

$$\begin{aligned}p(\hat{C}_{pK}) &\approx \mathcal{N}(\mu_C(\theta), \sigma_C^2(\theta)), \\ \mu_C(\theta) &= \frac{\mu_\mu - \text{LSL}}{3\mu_\sigma}, \\ \sigma_C^2(\theta) &= \left(\frac{\sigma_\sigma}{\sigma_\mu}\right)^2 \left(\left(\frac{\sigma_\mu}{3\sigma_\sigma}\right)^2 + \left(\frac{\mu_\mu - \text{LSL}}{3\mu_\sigma}\right)^2\right),\end{aligned} \tag{10}$$

which is valid for $\sigma_\sigma/\mu_\sigma \leq 0.1$. This value is just over the threshold for our considered application. We figured out that this is unproblematic through preliminary examinations so that we rely on the normal approximation.

3.2 Batch Constrained Bayesian Optimization

BO aims to find the global maximizer of an unknown function using an iterative procedure. Therefore, the method alternates between 1.) training the Gaussian process based on the currently available data 2.) solving an underlying optimization problem that depends on the trained Gaussian process and provides the next control parameter configuration that should be tested and 3.) testing the chosen parameters on the real system and obtaining a new observation that is added

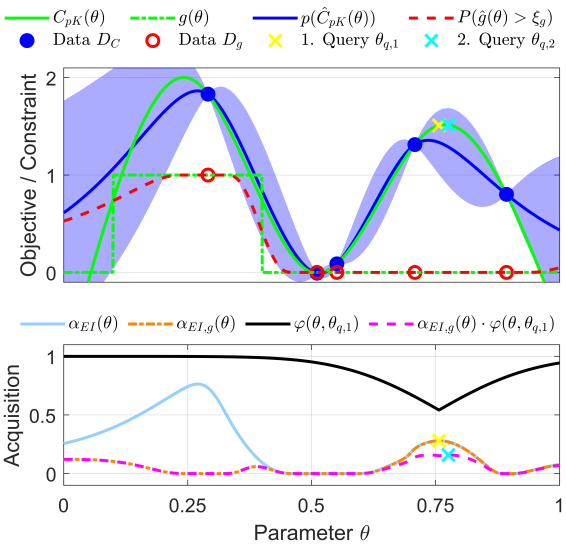

Figure 4: Illustrative representation for an iteration of our BO method in the one-dimensional case with one control parameter. A detailed explanation can be found in the text.

to the previous data. These steps are repeated until, e.g., a predetermined number of experiments have been performed. Then, the best control parameter-set with respect to (3) is selected from all observations collected so far.

In the following we present a detailed explanation of the execution of our batch constrained BO approach. For a better understanding of subsequent explanations, the reader is referred to Algorithm 1 and Figure 4. Algorithm 1 summarizes the steps of our BO implementation. In addition, Figure 4 shows the relationships of the considered functions during one iteration for a one-dimensional example. Note that the application to the bonding process is multi-dimensional. The assumption here is that we already have 5 evaluations of the real unknown functions. These are represented by the blue and red circles in the upper image. In addition, the true function $C_{pK}(\theta)$ for the process capability index and the constraint $g(\theta)$ are shown in solid and dashed green lines. With reference to the objective function, we also see the current Gaussian process assumption in blue. Here, the solid line is the mean and the shaded area is the standard deviation (in comparison to (7) and (10)).

A key component in step 2.) is the so-called acquisition function that defines the underlying optimization problem. The acquisition function is required to derive the new parameter configuration from the two sources of information available from the Gaussian process, i.e. the mean $\mu_C(\theta)$ and the uncertainty in form of the variance $\sigma_C^2(\theta)$. Here, we consistently use the criterion of expected improvement

$$\begin{aligned}\alpha_{EI}(\theta) &= \mathbb{E}[\max(0, \hat{C}_{pK}(\theta) - \xi_C)] \\ &= \int_{\xi_C}^{\infty} (\hat{C}_{pK}(\theta) - \xi_C)\mathcal{N}(\mu_C(\theta), \sigma_C^2(\theta))\, d\hat{C}_{pK} \\ &= \sigma_C(\theta)\phi(\gamma(\theta)) + (\xi_C - \mu_C(\theta))(1 - \Phi(\gamma(\theta))),\end{aligned}$$

$$\gamma(\theta) = \frac{\xi_C - \mu_C(\theta)}{\sigma_C(\theta)}, \tag{11}$$

where $\phi(\cdot)$ is the probability density function and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. The main idea here is to focus on the probability measure above a certain threshold ξ_C and use the center of this measure as a point to determine the next query location. The criterion thus strikes a balance between exploration and exploitation.

In the lower picture of Figure 4 we see the evaluation of the objective Gaussian process from the upper picture. The solid light blue line shows the expected improvement acquisition function from Equation (11). The associated threshold value ξ_C is 1. This is why the function is close to 0 in a range around the parameter value of 0.5. Maximizing this function results in the real functions from the upper image being evaluated near the left data point. This is the global maximum, however the constraint is not met in this region ($g(\theta) = 1$ for $\theta \in (0.1, 0.4)$). Thus, an evaluation is not desired at this location.

Next, we turn our consideration to the constraint for preventing tool collisions. Here the probability density distribution $p(\hat{g}(\theta) \mid D_g) = \mathcal{N}(\mu_g(\theta), \sigma_g^2(\theta))$ is given. In contrast to the objective function, we are not interested in the specific value for the constraint, but rather in the probability for the occurrence of a tool collision. For this reason, we integrate over the probability density function from a certain threshold ξ_g

$$\begin{aligned}P(\hat{g}(\theta) > \xi_g) &= \int_{\xi_g}^{\infty} \mathcal{N}(\mu_g(\theta), \sigma_g^2(\theta)) d\theta \\ &= \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{\theta - \mu_g}{\sqrt{2\sigma_g^2}}\right)\right),\end{aligned} \tag{12}$$

where $\operatorname{erf}(\cdot)$ is the so called error-function. We treat $g(\theta) = 0$ as a soft constraint and multiply the acquisition function (11) by the counter probability of (12), which yields

$$\alpha_{EI,g}(\theta) = \alpha_{EI}(\theta)\left(1 - P(\hat{g}(\theta) > \xi_g)\right). \tag{13}$$

In this way, even favorable regions where the probability for a tool collision is high and the acquisition function value is also high are still taken into account

Algorithm 1: Batch Constrained Bayesian Optimization.

Input: Initial dataset $D_1 = \{\theta_i, \mu_{F_S,i}, \sigma_{F_S,i}, g_i\}$, with $i = 1,...,n_{init}$, batchsize n_{batch}, iteration budget n_{budget}, mean functions m_μ, m_σ, m_g and covariance functions k_μ, k_σ, k_g, threshold values ξ_C, ξ_g, lower specification limit LSL.
for $i = 1$ **to** n_{budget} **do**
  Update Gaussian process hyper-parameters w.r.t. $\mu_{F_S}, \sigma_{F_S}, g_i$ based on D_i. ▷ By solving (6)
  for $b = 1$ **to** n_{batch} **do**
    Calculate b-th batch element $\theta_{q,b}$. ▷ By solving (15)
  end for
  Evaluate at $\theta_{q,b}$ and receive $\{\mu_{F_S,q,b}, \sigma_{F_S,q,b}, g_{q,b}\}$, for $b = 1,...,n_{batch}$.
  Attach new data to existing one $D_{i+1} = D_i \cup \{\theta_{q,b}, \mu_{F_S,q,b}, \sigma_{F_S,q,b}, g_{q,b}\}$.
end for
Returns: θ_r, with $r = \text{index max}_{r,g_r=0} C_{pK,r=1,...,n_{init}+n_{batch}n_{iter}}$.

and can thus be evaluated. This is especially important for the early iterations where a small amount of data is available and the Gaussian process with respect to the constraint provides poor predictions.

The red dashed line in the upper image of Figure 4 shows the probability of a tool collision (12). The threshold ξ_g was set to 0.5. It can be seen that the probability of a tool collision on the right side is broadly zero, whereas the probability changes smoothly to a value of one for the only observed tool collision on the left side. For a diminishing value θ towards zero, the probability drops again as we extrapolate here and have chosen an optimistic mean function of zero. The orange dashed line in the lower image shows the weighting of the expected improvement acquisition function with the counter probability for a tool collision (13). The left area is discounted down so that the new maximum is assumed to be on the right and should be evaluated at the location of the yellow cross.

Up to this point, we have only considered one experiment in each iteration. With respect to ultrasonic wire bonding, the first step is to insert a blank substrate plate into the bonding machine and automatically bond it with the selected control parameters. Then, the substrate plate must be removed from the bonding machine and placed into the shear tester, where the shear resistance is measured. The resulting values are then manually entered into the database on which the BO algorithm operates. Since bonding and shearing are relatively fast and automated, it is reasonable to calculate multiple control parametersets directly in one BO iteration instead of one and to evaluate them in parallel. These parallel evaluations are called batches and we will use the approach from (Gonzalez et al., 2016) for our implementation. To further motivate the usage of batches, we take a look at the specific time needed for one iteration. This time is composed of the time to calculate the query location t_{calc}, the time needed to prepare the experiment $t_{prepare}$ and the time spent for the actual evaluation t_{eval}. The total time then sums up to $T_{single} = (t_{calc} + t_{prepare} + t_{eval}) \cdot n_{iter}$, where n_{iter} is the number of iterations. For n_{batch} batch elements the total time amounts to $T_{batch} = (n_{batch} t_{calc} + t_{prepare} + n_{batch} t_{eval}) \cdot \frac{n_{iter}}{n_{batch}}$ under the assumption that the preparation time does not change and the time for calculation and evaluation scales linearly, which is a rather pessimistic assumption in most cases. For our experimental setting, the following times can be roughly estimated (in seconds) as $t_{calc} = 10, t_{prepare} = 100, t_{eval} = 50$. With $n_{iter} = 100$ and $n_{batch} = 6$, the total times correspond to $T_{single} = 4$ hours and $T_{batch} = 1.9$ hours, which is equivalent to a reduction of more than 50%. However, this reduction comes at the price of a decreased efficiency because all identified batch elements are based upon the same Gaussian processes and therefore, it is only updated after all batch elements are evaluated and not one by one.

To accommodate multiple query locations, the weighted acquisition function must be further modified. The approach in (Gonzalez et al., 2016) recommends the use of a so-called local penalizer

$$\varphi(\theta, \theta_q) = \frac{1}{2} \operatorname{erf}\Big(-\frac{1}{\sqrt{2\sigma_C^2(\theta_q)}} \big(\|G\|_2 \, \|\theta - \theta_q\|_2 - \xi_C + \mu_C(\theta_q) \big) \Big),$$
$$\text{with } G = \left. \frac{d\mu_C(\theta)}{d\theta} \right|_{\theta=\theta_q}. \tag{14}$$

This penalizer depends on the control parameters θ and on a specific query location θ_q, which is the previously selected batch element. The strength of the penalty depends on the norm of the gradient of the posterior mean function $\|G\|_2$ times the distance be-

tween the parameters under consideration and the previous batch element $\|\theta-\theta_q\|_2$. Other components are the posterior mean $\mu_C(\theta_q)$ and variance $\sigma_C^2(\theta_q)$ evaluated at the previous batch element and the threshold value ξ_C. The basic idea is to use a penalizer for each calculated batch element and determine the next batch element by the product of the soft constrained acquisition function and all local penalizers

$$\theta_{q,b} = \arg\max_{\theta} \begin{cases} \alpha_{EI,g}(\theta) & \text{if } b = 1, \\ \alpha_{EI,g}(\theta) \prod_{i=1}^{b-1} \varphi(\theta, \theta_{q,i}) & \text{else,} \end{cases} \tag{15}$$

for $b = 1, ..., n_{batch}$. In summary, the individual optimization problems (15) are solved sequentially one after the other until all n_{batch} batch-elements have been calculated. Afterwards, all batch elements respectively control parameters are tested on the real system. Figure 4 illustrates this concept by the black solid line representing the local penalizer (14) at the location of the yellow cross. The product of the black line and the orange dashed line gives the magenta dashed line, whose maximum value is at the position of the cyan cross (compare to (15)). For this illustrative example the total number of batch elements is two. Further batch elements can be calculated by additional local penalizers as required by the application.

4 APPLICATION

In order to obtain high quality results for the bond joints despite the complexity of the process in ultrasonic bonding, our proposal is to apply BO for determining the optimal feed-forward control. For this purpose, in this section we will first list our experimental settings regarding our specific equipment and chosen inputs for Algorithm 1 (Subsection 4.1). Then, the results from the real experiments are described and analyzed to clearly show the benefits in ultrasonic bonding (Subsection 4.2). To further strengthen the results, we also show the application of our method to a simulated bonding process, allowing us to study the behavior for many more initial conditions (Subsection 4.3).

4.1 Experimental Settings

For our experiments we used a Hesse Mechatronics automatic wire bonder BJ955 with an RBK03 bond head. Aluminum dibond plates are used as the substrate. The wire also consists of aluminum, has a diameter of 500 µm and is manufactured by Tanaka, type TANW Soft 2. Shear strengths are measured using a Xyztec Sigma shear tester.

With respect to Algorithm 1, we set the number of initial experiments n_{init} to 10, the number of batch elements n_{batch} to 6, the iteration budget n_{budget} to 15, which limits the number of experiments to $n_{init} + n_{batch} n_{budget} = 100$, and the threshold values ξ_C to 2 and ξ_g to 0.5. The lower specification limit LSL is 2500 cN and the number of bonds per control parameter-set n_{rep} is 10. The search space of the control parameters is box-constrained with lower limit θ_{lb} and upper limit θ_{ub} (see Table 1).

All Gaussian processes assume the 3/2-Matérn covariance function described in (5). In order to compare the hyper-parameters we also normalized the inputs to the unit interval via the lower and upper bounds from Table 1. Algorithm 1 was implemented in MATLAB. The optimization of the acquisition function is solved in two steps via random search with 1 million candidates. The best candidate is used as an initial guess for the follow up optimization with the build-in routine *fminsearch*[1]. Note that the massive number of evaluation of the acquisition function is unproblematic, since its computational complexity is low and it can be calculated relatively fast.

4.2 Real-world Results

In this subsection we present the results from the real bonding process. We compare 4 different approaches with each other:

- **Random Search:** A random controller parametrization is drawn from the uniform distribution $\theta_i \sim \mathcal{U}(\theta_{lb}, \theta_{ub})$ for every experiment.
- **Manual Tuning:** A non-expert, who is familiar with the process and its physical effects, is tuning the control parameters manually. The person has to choose 6 parametrizations in every iteration, which will then be evaluated in parallel. This is comparable to the proceeding of Algorithm 1. All previous experiments can be accessed at any time in order to draw conclusions for the next parametrizations.
- **BO with a Constant Prior Mean Function:** This is our base case scenario, where no specific a-priori knowledge about the process is available. Therefore, we set the prior mean functions for the mean shear force and the standard deviation to constant values, namely $m_\mu = 2500$ and $m_\sigma = 60$,

[1]More information can be found at mathworks.com/help/matlab/ref/fminsearch.html

Table 1: Box-constraints for the search space.

	F_0[cN]	F_1[cN]	F_2[cN]	$\hat{U}_1$[V]	$\hat{U}_2$[V]	T_1[ms]	T_1[ms]
θ_{ub}	300	375	375	43.3	7.75	5.5	29.5
θ_{lb}	900	1125	1125	80.6	54.25	38.5	206.5

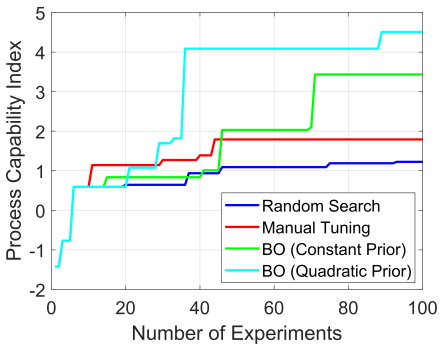

Figure 5: Progress of the objective over the number of performed experiments. An improvement is only achieved if a tested control parameterization leads to bonds with a higher process capability index than any previous parametrization and the optical criteria are satisfied. Otherwise the progress curve stays constant.

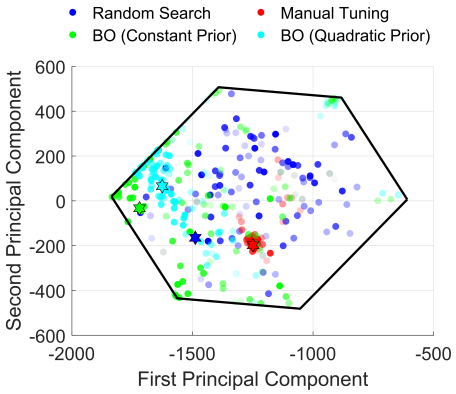

Figure 6: Progress in an reduced parameter space. The transparency indicates the number of the experiment, the greater the earlier. The stars show the best parameters for each approach.

resulting in a rather pessimistic initialization with a process capability index value of 0.

- **BO with a Quadratic Mean Prior Function:** This is a reference scenario, where a quadratic prior mean function $m_\mu(\theta) = \theta^T A\theta + b^T\theta + c$ for the mean shear force is used instead of a constant one. The mean function for the standard deviation stays constant. The quantities A, b, c are fitted to all data gathered from the other three approaches via least-squares regression.

In Figure 5, we see the progress in the process capability index over the number of experiments. Constant plateaus indicate that no improvement happened in the related experiments. Note that a jump in the objective only occurs if the C_{pK} value is higher compared to all previous experiments and the constraint, with respect to optical criteria, is satisfied. All approaches were given the same 10 initial experiments.

In industrial processes, a minimum C_{pK} of 1.33 to 1.67 is typically required (DVS - German Welding Society, 2017). Random search gets stuck near the value of 1 pretty fast and no significant improvement happens. From these observations, we suspect that the surface of the objective function is rather flat for a large part of the search space. Manual tuning provides an improvement in efficiency. However, after around 50 experiments, no further improvement was found. On the other hand, BO with a constant prior found a significantly better control, although it was previously on the same level as manual tuning. We think that the early random search like behavior is due to initial exploration of the search space. However, this behavior might be beneficial for later iterations. BO with a quadratic prior shows the best results and provides the best control over all approaches. Several experiments are needed to obtain this parametrization, which indicates that the quadratic prior has some mismatches to the real objective function. However, the exploration is guided to a region of potentially high quality parameters, which results in an overall high efficiency. This case shows how a reasonably good, but not necessarily perfect prior mean function improves the progress of BO. Instead of a data-driven model, we might think of a physical model of the process to improve performance. From a reverse-engineering perspective, our results show that this physical model has to imply a quadratic function for the shear force. We want to investigate this general idea in the future.

Next, we focus on the progress in the parameter space, see Figure 6. Since a 7-dimensional visualization of all parameters would be confusing, we transformed the data onto a 2-dimensional space by a principal component analysis (PCA). A singular value decomposition (SVD) of the data matrix, which is formed from all parameter values of all experiments, was calculated. Then, the parameter values were transformed into a 2-dimensional space via the first two columns of the matrix with the left singular vectors, which are thus assigned to the two largest singular values. The parameters related to random search are broadly diversified as expected. The concentration of red dots in one area is noticeable. This region was locally explored by the manual tuner, since the re-

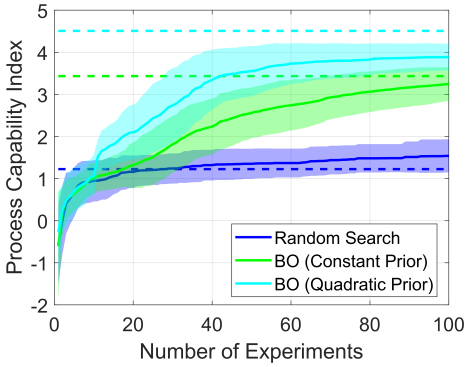

Figure 7: Progress of the objective function value over the number of experiments in the simulated environment. The solid lines show the average progress over multiple runs and the transparent areas indicate two standard deviations. The dashed lines show the maximum values found for the real process (compare to Figure 5).

sulting objective function values were relatively good and the constraint was fulfilled. We would classify the proceeding as safe exploration, where the region of tested parameters is iteratively expanded. We also noticed that the overview of the previous experiments was lost at around 30 to 40 experiments, which is the explanation for the lack of progress on the objective function in Figure 5. On the other hand, BO especially explores the bounds of the search space and focuses relatively fast on the region where the highest values were found. For reasons of confidentiality, we are not permitted to provide the optimal parameters.

4.3 Simulated Results

Since experiments on the real system are time and cost consuming, we could not investigate the influence of the initial experiments and therefore the robustness of each approach. Based on this reason, we set up Gaussian processes, which are built upon all data gathered so far, to replace the real system. Hence, their predictions are relatively accurate. This virtual environment enables us to test the approaches used in Subsection 4.2, excluding manual tuning, for different initial experiments.

Figure 7 shows the dedicated results for 50 different runs. First of all, we see that the runs with respect to the real system are reasonable, because they fit the simulated runs. Furthermore, the used BO methods outperform random search robustly and converge to high objective function values. The fact that the global maximum from the measurements (dashed cyan line) is not reached by quadratic prior BO may be related to the regularization of the reference Gaussian processes. However, the optimal parameters found are identical. These results strengthen our hypothesis that BO is appropriate for the control design of the bonding process.

Another advantageous property of training Gaussian processes with all data is the ability to investigate the learned hyper-parameters η_*. Especially the learned scales reveal the relevance of a given parameter dimension to the output. This is called automatic relevance determination in the literature (Rasmussen and Williams, 2006).

The hyper-parameters are shown in Table 2. Note that a direct comparison is possible, since we standardized the parameter dimensions to the unit interval. The higher the value of the scale l, the smaller the influence of the underlying parameter. Regarding the mean shear force $\hat{\mu}_{F_S}(\theta)$ and the label $\hat{g}$, we see that the normal force of the pre-deformation phase F_0 is less relevant compared to the standard deviation $\hat{\sigma}_{F_S}(\theta)$, which seems plausible, because a sufficient initial contact area is formed by a wide range of values. On the other hand, the values of the ultrasonic voltages $(\hat{U}_1, \hat{U}_2)$ have a relatively high impact on the mean shear force. This also holds for the standard deviation of the shear force $\hat{\sigma}_{F_S}(\theta)$. However, we see that the force F_2 and time T_2 of the second phase have the most influence, which means that these values have to be chosen with high accuracy for receiving a low standard deviation. The values for the signal and noise variance (σ_f^2, σ_n^2) match with the observations during the experiments.

5 CONCLUSION AND OUTLOOK

Real-world and simulated results showed that the application of BO to the feed-forward control design of ultrasonic wire bonding is very appropriate. The suggested approach outperforms random search and manual tuning in terms of efficiency and robustness. We also showed that the incorporation of a quadratic prior mean function is advantageous and increases the performance. We see this result as a guideline for further research, where we exchange the quadratic prior with a physical simulation model of the bonding process. Our work lays the foundation for this since it builds upon numerous measurements with a wide range of control inputs applied to the real bonding process and is thus well validated. Furthermore, we investigated the hyper-parameters and discovered that the touchdown force, which is applied during the pre-deformation phase, has little influence on the bond quality. The mean shear force is sensitive to the voltage amplitude values, whereas the variance responds the most to the normal force in the second phase and its duration.

Table 2: Hyper-parameters for the reference Gaussian processes.

$\mathcal{GP}$	σ_f^2	l_{F_0}	l_{F_1}	l_{F_2}	l_{U_1}	l_{U_2}	l_{T_1}	l_{T_2}	σ_n^2
$\hat{\mu}_{F_S}(\theta)$	$4.4 \cdot 10^5$	13.7	2	1.9	1.7	1.4	2.9	2.6	$1.6 \cdot 10^3$
$\hat{\sigma}_{F_S}(\theta)$	$1.7 \cdot 10^3$	1.3	1.9	0.2	0.9	0.9	2.3	0.8	335
$\hat{g}$	0.3	3	0.7	1.1	0.5	0.5	1	1.3	0.05

Besides a physics-based prior mean function, we want to investigate other materials, like copper, or different wire diameters. The modification of other general conditions could also be considered. Our proposed method can be applied in the same way. However, we can build on the results from this paper and examine the field of transfer learning. A good starting point might be the Gaussian processes which were trained with all the data points from our experiments.

ACKNOWLEDGEMENTS

The research was funded by the Ministry of Economic Affairs, Innovation, Digitalisation and Energy (MWIDE) of the State of North Rhine-Westphalia within the Leading-Edge Cluster Intelligent Technical Systems OstWestfalenLippe (it's OWL) and by the Federal Ministry of Education and Research of Germany (BMBF) within the junior research group DART of the University of Paderborn. The responsibility for the content of this publication lies with the authors.

The authors would like to thank Yuqi Liu, Jan Herbermann and Fabian Reiling for their assistance during the experiments.

REFERENCES

Calandra, R., Seyfarth, A., Peters, J., and Deisenroth, M. P. (2016). Bayesian Optimization for Learning Gaits under Uncertainty. *Annals in Mathematics and Artificial Intelligence*, 76:5–23.

Díaz-Francés, E. and Rubio, F. J. (2013). On the Existence of a Normal Approximation to the Distribution of the Ratio of two Independent Normal Random Variables. *Statistical Papers*, 54(2):309–323.

DVS - German Welding Society (2017). Test Procedures for Wire Bonded Joints (Technical Bulletin DVS 2811).

Geißler, U. (2009). Verbindungsbildung und Gefügeentwicklung beim Ultraschall-Wedge-Wedge-Bonden von AlSi1-Draht. *Technische Universität Berlin, Fakultät IV - Elektrotechnik und Informatik*.

Gogh, B., Benner, T., Seppaenen, H., Tszeng, C., and Sepehrband, P. (2020). An Oxide Wear Model of Ultrasonic Bonding. *International Symposium on Microelectronics*, 2020:222–229.

Gonzalez, J., Dai, Z., Hennig, P., and Lawrence, N. (2016). Batch Bayesian Optimization via Local Penalization. In *AISTATS 2016; 19th International Conference on Artificial Intelligence and Statistics*.

Harman, G. (2010). Wire Bonding in Microelectronics. McGraw-Hill.

Hunstig, M., Schaermann, W., Broekelmann, M., Holtkaemper, S., Siepe, D., and Hesse, H. J. (2020). Smart Ultrasonic Welding in Power Electronics Packaging. In *CIPS 2020; 11th International Conference on Integrated Power Electronics Systems*, pages 1–6.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13(4):455–492.

Kelly, M. (2017). An Introduction to Trajectory Optimization: How to Do Your Own Direct Collocation. *SIAM Review*, 59(4):849–904.

Long, Y., Twiefel, J., and Wallaschek, J. (2017). A Review on the Mechanisms of Ultrasonic Wedge-Wedge Bonding. *Journal of Materials Processing Technology*, 245:241–258.

Mayer, M. and Schwizer, J. (2002). Ultrasonic Bonding: Understanding How Process Parameters Determine the Strength of Au-Al Bonds. Symposium on Microelectronics, IMAPS, Denver.

Neumann-Brosig, M., Marco, A., Schwarzmann, D., and Trimpe, J. S. (2020). Data-efficient Auto-tuning with Bayesian Optimization: An Industrial Control Study. *IEEE Transactions on Control Systems Technology*, 28(5):730–740.

Rasmussen, C. E. and Williams, C. (2006). Gaussian Processes for Machine Learning. MIT Press.

Schemmel, R., Krieger, V., Hemsel, T., and Sextro, W. (2020). Co-Simulation of MATLAB and ANSYS for Ultrasonic Wire Bonding Process Optimization. In *2020 21st International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems*.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N. (2016). Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, 104(1):148–175.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Unger, A., Hunstig, M., Meyer, T., Brökelmann, M., and Sextro, W. (2018). Intelligent Production of Wire Bonds using Multi-Objective Optimization – Insights,

Opportunities and Challenges. *International Symposium on Microelectronics*, 2018(1):572–577.

Unger, A., Sextro, W., Althoff, S., Meyer, T., Brökelmann, M., Neumann, K., Reimann, R. F., Guth, K., and Bolowski, D. (2014). Data-driven Modeling of the Ultrasonic Softening Effect for Robust Copper Wire Bonding. In *Proceedings of 8th International Conference on Integrated Power Electronic Systems*, volume 141, page 175–180.