Surrogate-based Evolutionary Optimization for Friction Stir Welding

Cem C. Tutum
University of Texas at Austin
Department of Computer Science
78712 Austin, TX, USA
Email: tutum@cs.utexas.edu

Shaayaan Sayed
University of Texas at Austin
Department of Computer Science
78712 Austin, TX, USA
Email: ssayed@cs.utexas.edu

Risto Miikkulainen
University of Texas at Austin
Department of Computer Science
78712 Austin, TX, USA
Email: risto@cs.utexas.edu

***Abstract*—Friction Stir Welding (FSW) is an innovative manufacturing process, which is used to join two pieces of metal with frictional heating and plastic deformation due to stirring action. Melting is avoided during the process, therefore problems related to microstructure phase transformation (i.e., cooling from the liquid phase) are avoided. The temperature distribution in the weld zone, as a function of the heat generation, highly affects the evolution of the residual stresses in the work piece, hence the performance of the final product. Therefore, thermal models play a crucial role in detailed analysis and improvement of this process. In this study, a previously developed and validated three-dimensional steady state thermal model of FS welding of AA2024-T3 plates has been used for evaluating the quality of the candidate solutions. It should be noted that this is a computationally expensive model and closed form formulations (i.e. analytical equations) for the underlying physics are not available, which forces us to use them sparingly during the optimization procedure. A mathematical correlation model, a *surrogate* in other words, is iteratively constructed to replace the FSW simulations and guide the search towards feasible and promising regions. A new surrogate-based optimization algorithm named EICTS, Expected Improvement with Constrained Tournament Selection has been developed. The striking difference of EICTS from other surrogate based constrained optimization methodologies that it needs to construct only two surrogates, i.e. one for the objective function and another one to handle all constraint functions (i.e., instead of approximating each of them individually). EICTS is first tested on some well-known engineering problems with multiple constraints and finally on the FSW problem briefly mentioned above. Its *runtime* and *convergence* performances are compared with EIPF (Expected Improvement with Probability of Feasibility) method and found very promising.**

I. INTRODUCTION

The friction stir welding (FSW) process is getting more attractive especially in aerospace and automotive industries where there is a high demand for lightweight structures built of materials having high strength-to-weight ratio such as aluminum alloys [1], [2]. First and foremost, the mechanical properties of the metal are preserved as much as possible since there is no melting during the process. It is also advantageous in case of welding large structures which cannot be heat-treated afterwards. The process starts with clamping the workpieces on to an anvil to avoid abutting surfaces spread apart. Then a rotating wear-resistant tool is submerged and traversed along the joint line while stirring the two pieces of metal together. The frictional heating, together with the plastic work provided by the forging and stirring motions, softens the material and makes the FSW tool move forward easier. The process is finalized by removing the tool out of the two workpieces and by let- ting it cool down to form the weld. These steps have been schematically shown in Fig. 1.

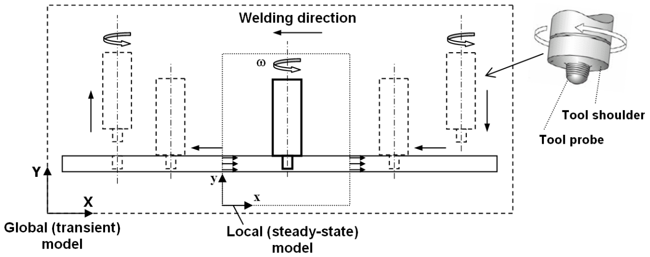

Fig. 1. Schematic view of the FSW process and a typical FSW tool.

Since the development of the process in 1991, FSW modeling studies involving several research areas such as heat transfer [3], [4], [5], [6], material flow [7], material science and metallurgy [8], and solid mechanics [3], [4], [9], [10], [11] are increasing every year. However, the common ground behind those models is the requirement for high demanding computation time. Therefore, the number of numerical optimization studies is limited [12], [13], and design improvements mostly were performed by experimental works. Most of these numerical optimization studies are based on pure thermal models concerning single objectives or serving for inverse modeling purposes. The reason why the emphasis is put on thermal models is not only the less requirement for the computational resources, but also the dominant relationship between the microstructure as well as the residual stress evolution in the weld zone and the final the performance of the weld. Therefore, thermal models play a crucial role in detailed analysis and improvement of this process. Few multiobjective optimization studies regarding thermal [14], [15] and thermo-mechanical aspects (i.e., residual stresses) of the FSW process have recently been presented by [16].

Having given an overview about the general activities and challenges about the multi-physics simulation and optimization of the FSW process [12], [13], it is worth here to mention about the computationally more efficient optimization algorithms which are based on approximating methods (*surrogates*, meta-models, or response surfaces in mathematical parlance [17], [18], [19], [20], [21], [22], [23], [24]). Most known surrogates in the literature vary from simple polynomial regression models and moving least squares to neural networks, radial basis functions, Kriging, and support vector regression. Some of these methods are also listed under machine learning, statistical learning, or in general supervised learning techniques. Despite the variety in their mathematical construction, they all work on the same consecutive principles: *training* (learning) and *testing* (prediction or generalization). Training, in simple terms, is the procedure of learning the behavior of the underlying response as a function of some chosen parameters, which can also be called as the mathematical mapping. Once this mapping is learned using limited sample size, e.g., in case of using multidisciplinary manufacturing process simulations, it can be used to iteratively replace the computationally expensive black-box function. This simpler model allows the user to predict any response at an unknown design set at a negligible cost.

This article is structured as follows. First, a three-dimensional steady state thermal model of FSW of AA2024-T3 plates simulated in COMSOL has been presented in section II. Next, a brief introduction to Kriging, i.e,. the surrogate (approximation function), to be used in the Efficient Global Optimization (EGO) (*see* (section III-B)) is described in section III-A. Then two update criteria, EIPF and the proposed method EICTS, are introduced in sections III-C and III-D respectively. The results of the validation cases as well as the simulation-based optimization problem in FSW are given in section IV. Following the discussions of the results, the article is finalized with the concluding remarks and addressing the future work.

II. SIMULATION OF FSW

In this study, the temperature distribution during FSW of two AA2024-T3 plates has been simulated using commercial multi-physics finite element software COMSOL [25]. The plunge, the dwell, and the pull-out periods have been neglected; therefore, only the steady-state period has been the focus of the modeling procedure. This assumption alleviates some remarkable complexities and still allows capturing the first-order effects the process parameters on the fully developed temperature field in the welding domain [5], [6]. Moreover, the performance of the numerical optimization is enhanced by reducing the computational cost of the model, hence increasing the number of function evaluations in a limited time frame [14], [12].

The heat generation in FSW is often simulated by the application of a surface heat flux in case of sliding or a volume heat flux in case of sticking boundary condition [5], [6]. However, the uncertainty in the contact status at the interface reveals other problems related to the model input parameters such as the friction coefficient or the yield shear strength. In most of the studies, common effort is put into deriving the total heat generation or the friction coefficient from the tool force and torque measurements as the main heat source. However, this is conflicting with the aim of the thermal model that is in essence used to predict the heat generation. Besides, the need for the measurements for each different case having different process parameters would not be straightforward. In order to overcome these limitations, thermal models should be integrated with mechanical and flow models in which contact condition, deformation, etc. can be captured simultaneously. On the other hand, taking all these effects into account would obviously be an overkill for an optimization study since even only one function evaluation would take several days or weeks to compute [12].

In the present optimization study, a three-dimensional steady-state Eulerian thermal model involving an analytical heat source called Thermal-Pseudo-Mechanical (TPM) model [6], which is capable of incorporating with the prescribed material flow [5], has been implemented in COMSOL. It is a well-known phenomenon that once the material is heated, its yield stress decreases and vice versa. When the temperature exceeds the solidus temperature, it behaves as a fluid having almost no resistance to deformation; therefore, contribution to heat generation becomes negligible. Therefore, the solidus temperature acts as a switch button for turning the heat generation from "on" to "off". This information has been implemented here in the TPM model. As evident from its name, it is a purely thermal model which takes some mechanical information (i.e., yield strength variation of the metal with temperature) into account, thus bypassing the solution of the mechanical field. Even though some nonlinearity is introduced by application of this solution (i.e., temperature) dependent heat source, the computation time is still much more reasonable as compared to the solution time of a full thermo-mechanical model. The details of the model, i.e., heat source, boundary conditions, enmeshment, etc., are given in the following paragraphs.

For the purpose of predicting the thermal field in the workpiece, the classical time-dependent heat conduction equation should be solved regardless of the complexity of the heat source, see Eq. 1,

$$\rho c_p \dot{T} = \nabla(k \nabla T) + q_{vol}, \tag{1}$$

where ρ (kg/m^3) shows the material density, c_p (J/kgK) the specific heat capacity, T (K) the temperature output, k (J/mK) the heat conductivity, and q_{vol} (W/m^3) the volumetric heat source term. In case of describing the heat flow in a Eulerian reference frame under steady-state conditions, the time dependent term is removed and a convective term is added to Eq. 1 as given by the following equation,

$$0 = \nabla(k \nabla T) + q_{vol} - \rho c_p u \dot{T}, \tag{2}$$

where u is the velocity field vector defined analytically in the shear layer region (around the tool) which changes between the tool welding velocity and the tool rotational speed [5]. The heat generation (q_{vol}) defined in the shear layer regions is varying with the rotational speed (ω), the position $(r(x, y))$,

TABLE I
THERMO-PHYSICAL AND MECHANICAL PROPERTIES FOR THE AA2024-T3 [26]

T [$°C$]	k [W/mK]	ρ [kg/m^3]	c_p [J/kgK]	σ_y [MPa]
20	100	2780	929	306
100	121	2767	969	261
200	136	2749	1022	152
300	137	2729	1075	57
400	124	2709	1128	13
500	98	2689	1181	5

the yield strength of AA2024-T3 ($\sigma_{yield}(T)$), and the shear layer thicknesses (t_{SL}), see Eq. 3:

$$q_{vol}(x, y, T) = \frac{\omega r(x, y)\sigma_{yield}(T)}{\sqrt{3}t_{SL}}, \tag{3}$$

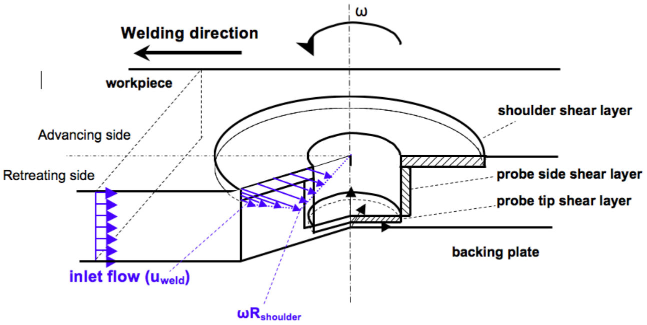

Fig. 2. Temperature distribution for the steady-state thermal FSW model. The figure on the left side is adapted from [5].

The calculation domain, shown in Fig. 3, composed of a 7 mm-thick plate geometry. The volume of the tool pin is removed. The volume heat flux given in Eq. 3 is defined in the cylindrical volumes drawn around the tool (Fig. 2). Temperature dependent AA2024-T3 thermo-physical material properties are assigned in the calculation domain (see Table I). The prescribed velocity field is implemented in the same way as in [5], the details are given in the original study. The velocity vectors (u and v) in the shear layers can be prescribed as in Eq. 4,

$$\begin{aligned} u &= -y\omega(1-\zeta) + \zeta u_{weld}, \\ v &= x\omega(1-\zeta), \end{aligned} \tag{4}$$

where u_{weld} is the tool feed rate and ζ is the position dependent interpolation coefficient in a shear layer region, i.e.,

$$\begin{aligned} \zeta_{Sh-side}(z) &= \frac{t_{plate}-z}{t_{Sh-side-SL}}, \\ \zeta_{Pr-side}(r) &= \frac{t-R_{probe}}{t_{Pr-side-SL}}, \\ \zeta_{Pr-tip}(z) &= \frac{z_{Pr-tip}-z}{t_{Pr-tip-SL}}, \end{aligned} \tag{5}$$

where t_{plate}, $t_{Sh-side-SL}$, $t_{Pr-side-SL}$, and $t_{Pr-tip-SL}$ are, respectively, the thicknesses of the plate, the shoulder side, the probe side as well as the probe tip shear layers, and z_{Pr-tip} is the vertical coordinate of the probe tip.

Fig. 3 represents a characteristic temperature distribution in an FSW application with the appropriate thermal boundary conditions. The front surface of the plate is kept constant at the room temperature (20 $°C$). The convection heat flow is applied at the back surface of the workpiece. Thermal insulation is enforced on both sides of the plate. Asymmetric temperature distribution (i.e., higher temperature values are observed at the advancing side over the retreating side) caused by the material flow close to the tool shoulder is also captured as very well known [2], [27]. The anvil and the FSW tool are neglected in this model. This is due to two reasons: first, inclusion of these will introduce extra uncertainties in the model (temperature boundary condition at the tool holder head, the heat flux diffused into the tool, etc.), and second, the size of the enmeshment will get larger, and this will immensely increase the computation time. The steel backing plate is equivalently replaced by a heat flux boundary condition having a heat transfer coefficient (HTC) of 1000 W/m^2K [27]. The cooling from the top surface is also taken into account by applying an HTC value of 10 W/m^2K.

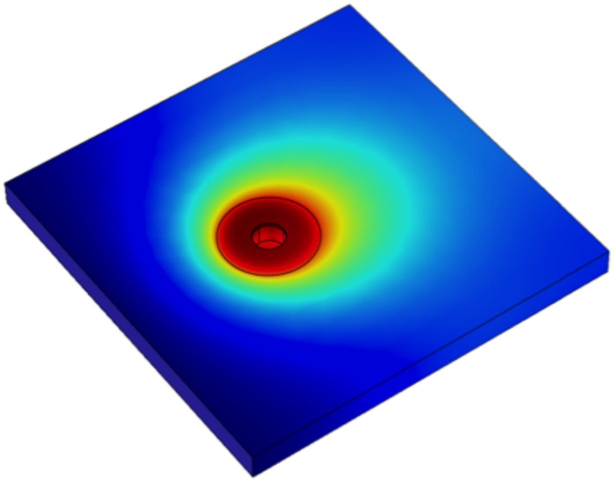

Fig. 3. Temperature distribution for the steady-state thermal FSW model.

III. OPTIMIZATION

A. Surrogate Model: Kriging

Kriging is a well-known surrogate technique that is frequently used to approximate computationally expensive functions in the course of optimization. The method, named after a South African geologist D. Krige [28], was developed to estimate mineral concentrations within a particular field and popularized by the work of Sacks et al. [29], which made it also known as Design and Analysis of Computer Experiments. The procedure starts with obtaining a sample data of limited size (i.e., n-design sets each having d-variables), $\mathbf{X_{n\times d}} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(n)}]^T$, and a corresponding vector of scalar responses $\mathbf{y_{n\times 1}} = [y^{(1)}, y^{(2)}, ..., y^{(n)}]^T$. It is assumed that if design points, e.g., $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$, are positioned close together in the design space, their respective function values $y^{(i)}$ and $y^{(j)}$ are expected to be similar, and vice versa. This can be formulated statistically by considering the correlation between two points as,

$$cor\left[y(\mathbf{x}^{(i)}),y(\mathbf{x}^{(j)})\right]=\prod_{k=1}^{d}exp\left(-\theta_k\left|\mathbf{x}_k^{(i)}-\mathbf{x}_k^{(j)}\right|^2\right),\quad (6)$$

where θ_k is a correlation parameter or hyperparameter (i.e., $\theta_k = \theta_1, \theta_2, ..., \theta_d$) which controls how fast the correlation changes from one point to the other one along each dimension. Here, Gaussian basis function is used; therefore the exponent is fixed at 2 yielding a smooth and continuous transition at $\mathbf{x}^{(i)}$. Eq. 6 is used to build the symmetric correlation matrix ($\mathbf{R}$) of all n-points in $\mathbf{X}$, which will be used in the process of tuning the unknown hyperparameter θ_k to maximize the likelihood of the assumed Gaussian model on the given dataset. Having the Kriging model parameters tuned, the next step is to predict a new response value, i.e., an objective or a constraint function value, at an unobserved design point using the sample data that are used to train the Kriging model. Ordinary Kriging predictor ($\hat{y}$) has such a form,

$$\hat{y}(\mathbf{x}^*)=\hat{\mu}+r(\mathbf{x}^*,\mathbf{x})^T\mathbf{R}(\mathbf{x})^{-1}\left(y(\mathbf{x})-1\hat{\mu}\right),\quad (7)$$

where $r(\mathbf{x},\mathbf{x})$ is the linear vector of correlations between the unknown point to be predicted ($\mathbf{x}^*$) and the known sample points ($\mathbf{x}$), ($\hat{\mu}$) is the estimated mean, and 1 is an unit vector of size n x 1. *Ordinary Kriging* assumes a constant term ($\hat{\mu}$) for the global fitting term in the predictor equation, whereas the *Universal Kriging* uses a known functional form. The second part on the right side of Eq. 7 represents the local deviation from the global term. Kriging is in general known for its good performance in fitting complex functional behavior; however, what makes Kriging a very popular surrogate technique is in essence its ability to estimate the *mean squared error* (MSE) at the unknown point,

$$\hat{s}^2(\mathbf{x}^*)=\hat{\sigma}^2\left[1-\mathbf{r}^T\mathbf{R}^{-1}\mathbf{r}+\frac{1-\mathbf{1}^T\mathbf{R}^{-1}\mathbf{r}}{\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}}\right],\quad (8)$$

where $\hat{s}^2$ represents the MSE estimate. The third term inside the square parentheses is very small and is often neglected. Since Kriging is an interpolation method, $\hat{s}^2$ reduces to zero at the sample points and consequently $\hat{y}$ becomes equal to the corresponding response value.

B. Efficient Global Optimization (EGO)

Knowing the fact that the Kriging model just constructed on the limited number of sample points (*initial sample set*) is only an approximation for the underlying black-box function; thus, new sample points (*infill points*) should iteratively be sought to update or in other words to improve the accuracy of the surrogate. This update procedure, i.e., *infill criterion*, may consider either only focusing on the optimum region of the predictor (i.e., running the risk of premature convergence) or to continue exploring the search space to increase the overall accuracy thereby having a higher probability of finding the global optimum. Another strategy is to balance both efforts, i.e., simultaneously utilizing the information of the predictor $\hat{y}$(x) calculated by Eq. 7 and the estimation of the variance $\hat{s}^2$(x) calculated by Eq. 8. Jones et al. [30] suggested an algorithm called *Efficient Global Optimization* (EGO), which relies on building iteratively a probabilistic model (i.e., *Kriging*, section III-A) of the objective function and a criterion based on improving upon the best sample found so far, y_{best}, by searching this probabilistic model. Recall that the Kriging predictor is the realization of a Gaussian process Y($\mathbf{x}$) with the mean $\hat{y}$ and the variance $\hat{s}^2(\mathbf{x})$; therefore, due to the uncertainty in the predictor, an improvement at a point $\mathbf{x}$ can be defined as,

$$I(\mathbf{x})=max\left(y_{best}-Y(\mathbf{x})\right),\quad (9)$$

which can be used to maximize the expectation of it (*expected improvement*) as the infill criterion ([31], [32]),

$$\begin{aligned}E[I(\mathbf{x}])=&(y_{best}-\hat{y}(\mathbf{x}))\Phi\left(\frac{y_{best}-\hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right)+\\&\hat{s}(\mathbf{x})\phi\left(\frac{y_{best}-\hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right),\end{aligned}\quad (10)$$

where $\Phi(.)$ and $\phi(.)$ are the *cumulative distribution function* and the *probability density function* of a normal distribution, respectively. Readers are referred to [33] for the derivation of Eq. 10. EGO iterates until a user-defined stopping criterion is met, e.g., total number of infill points, change in the objective function, tolerance on MSE, etc.

C. EIPF

As berifly described in the previous section, EGO framework is initially developed to handle unconstrained optimization problems by maximizing the expected improvement of the objective function. Schonlau [34] had suggested an intuitive and effective methodology which modifies the expected improvement idea in a way to handle the constraint functions as well. The idea is simply to convert the constrained optimization problem into an unconstrained one by multiplying the standard expected improvement value with the probability value of that point being feasible. The Probability of feasibility of a solution for a single constraint could be calculated as in Eq. 11,

$$P[F(\mathbf{x})]=P[g\leq g_{limit}]=\int_{-\infty}^{g_{limit}}\phi(g)dg,\quad (11)$$

where g_{limit} is the constraint limit. This new constrained expected improvement criterion, which is called as *EIPF* in short,

$$E[I(\mathbf{x})\cap F(\mathbf{x})]=E[I(\mathbf{x})]P[F(\mathbf{x})],\quad (12)$$

is maximized like in the unconstrained case. Hence, if the point to be evaluated is located in a feasible region (i.e., $P[F(\mathbf{x})] \rightarrow 1$), the constrained expected improvement value would be equal to $E[I(\mathbf{x})]$, otherwise if the solution is estimated to be in the infeasible region (i.e., $P[F(\mathbf{x})] \rightarrow 0$), then the constrained expected improvement value would be zero. EIPF criterion is implemented in the EGO framework and used as infill criterion.

TABLE II
A SAMPLE OF 8 SOLUTIONS WITH THEIR OBJECTIVE (Y_i) VALUE AND CONSTRAINT VIOLATION (CV_i^{1-3}) VALUES

$\mathbf{X_i}$	$\mathbf{Y_i}$	$\mathbf{CV_i^1}$	$\mathbf{CV_i^2}$	$\mathbf{CV_i^3}$
X_1	3628.2	0.10786	0.048429	13.129
X_2	2144.4	-0.72429	-0.43514	-13.994
X_3	39.876	-0.42857	-0.85714	14.075
X_4	151.34	0.12786	0.061571	14.042
X_5	172.75	-0.005	-0.21757	14.048
X_6	2332.8	-0.82714	-0.30543	-12.829
X_7	3360.3	-0.025	-0.34071	-12.944
X_8	4323.5	-0.02	-0.72743	13.276

TABLE III
CONSTRAINT VIOLATION (CV_i^{1-3}) VALUES ARE NORMALIZED. ROWS WHICH HAVE *only* ALL NEGATIVE OR ALL POSITIVE VALUES ARE SUMMED UP, WHEREAS IN OTHER ROWS, ONLY POSITIVE VALUES ARE SUMMED UP.

$\mathbf{X_i}$	$\mathbf{Y_i}$	$\mathbf{CV_i^1}$	$\mathbf{CV_i^2}$	$\mathbf{CV_i^3}$	$\mathbf{TCV_i}$
X_1	3628.2	10^{-8}	10^{-8}	10^{-8}	3x10^{-8}
X_2	2144.4	-0.8748	-0.3402	-1.0	-2.2151
X_3	39.876	-0.5152	-1.0	1.0	1.0
X_4	151.34	1.0	1.0	0.9658	2.9658
X_5	172.75	0.0	0.0	0.9715	0.9715
X_6	2332.8	-1.0	-0.1374	0.0	-1.1374
X_7	3360.3	-0.0243	-0.1925	-0.0988	-0.3157
X_8	4323.5	-0.0182	-0.7972	0.1554	0.1554

D. *Proposed Method:* ***EICTS***

In this paper, a new constraint handling methodology is proposed within the EGO framework without transforming the constrained problem into an unconstrained one. As usual, a uniform sampling method such as optimal *Latin Hypercube Sampling* (LHS) method is applied to obtain the initial sampling set within the bounded design space. Next, the true objective and constraint function values (i.e. high-fidelity simulations) are computed at these initial design sites. Up to this point, it is same with the standard EGO (both with $E[I(x)]$ and $EIPF$ criterions) algorithm. The proposed method needs to build only two surrogates; one for the objective function and another one for the total constraint violation (TCV), which is computed using all constraint violations. The constraints are defined in a way that negative values indicate that the solution is $feasible$ whereas the positive value indicates that the solution is $infeasible$. While computing the TCV value, only positive values (i.e. CV_i values of those infeasible solutions) are taken into account. However, building only one surrogate for multiple surrogates having values at different orders of magnitudes requires an additional normalizing step. To make the following discussion more clear, a sample of 8 solutions having an objective (Y_i) and three constraint violation values (CV_i^{1-3}) is provided in Table II,

In the next step, each CV_i column is checked separately; feasible solutions ($CV_i \leq 0.0$) are normalized between -1.0 and 0.0, infeasible solutions ($CV_i > 0.0$) are normalized between 10^{-8} and 1.0. Then, in order to compute single TCV_i value for each X_i solution, rows having only positive or only negative values are summed up, however in other rows, only positive values are summed up. The normalized CV_i and the computed TCV_i values are shown in Table III,

According to Table III solutions X_1, X_3, X_4, X_5, X_8 are infeasible and the other 3 solutions are feasible. These TCV_i values are then approximated using Kriging function. The next important step is to compute the standard $E[I(x)]$ values for all candidate solutions in the course of global search algorithm. The purpose is to find feasible candidate solutions (i.e., $TCV_i \leq 0$) with maximum $E[I(x)]$ values. To efficiently search the constrained design space *Constrained Tournament Selection*, with tournament size equal to 5, is implemented. During the iterations of the global optimization algorithm, 5 randomly selected individuals are first compared with respect to their TCV_i values: **i)** if all (or at least one) TCV_i values are negative (i.e., feasible) then the solution with the maximum $E[I(x)]$ value wins the tournament, **ii)** if all TCV_i values are positive (i.e., infeasible) then the solution with the minimum TCV_i value wins the tournament. At the end of the global optimization procedure, the optimal candidate solution, which is found using the surrogate models, is re-evaluated using high-fidelity simulations.

IV. RESULTS

In the following sections, three well known analytical engineering constrained optimization problems are given to validate the performance of the proposed method, EICTS. For each test problem, 30 experiments with different initial LHS set are performed. These three problems, *Gas Transmission Compressor Design* (GTCD), *Pressure Vessel Design* (PVD), *Welded Beam* (WB), are well studied by the constrained optimization researchers, however the application of surrogate-based constrained optimization algorithms are limited. Moreover, the performance of EICTS is only compared with EIPF due to the limited availability of the alternative open-source algorithms. Our main objective was to achieve comparable results in shorter computational time. The summary of the results in Table IV as well as in Figs. 4, 5, 6 and 7 show that even better solutions (i.e., better feasible objective values as well as closer solutions to the analytical optimum solution) are obtained in shorter time period.

A. *Validation Cases*

In the following three engineering problems, the global optimum solutions are known and reported in the literature. **30** experiments are performed for each validation case using both EIPF and EICTS methods. All problems have four design variables. The initial sample sets include **20** solutions and a *budget* of **40** infill point calculations are allowed for each experiment. Normalized distance of the *current best* solution to the analytical optimal solution and the 95% confidence intervals are shown in the following figures. Since algorithms are tested with the same initial sample sets, the average best and variance values between 1^{st} and 20^{th} samples are same. The infill points start to deviate from each other due to different algorithm performances.

1) Gas Transmission Compressor Design (GTCD): The problem is taken from [35]. It has 4 design variables and 1 constraint function. It is clearly seen from Fig. 4 that EICTS is able to find better solutions (i.e., closer to optimal solution).

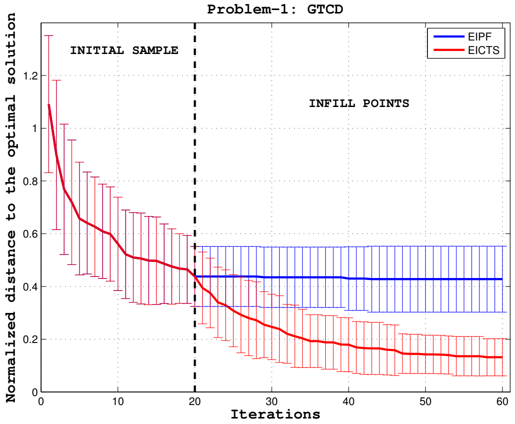

Fig. 4. Average of the smallest Euclidean distance to the optimal solution in 30 experiments for the GTCD problem.

2) Pressure Vessel Design (PVD): The problem is taken from [36] and [37]. It has 4 design variables and 3 constraint functions. In this case, the performances look closer but EICTS still performs better. The runtime performance of the two algorithms are also given in Table IV and EICTS starts to get faster as compared to EIPF, because different surrogate approximations for the multiple constraint functions need to be performed by EIPF, whereas only single surrogate approximation for all the constraint functions is needed for EICTS.

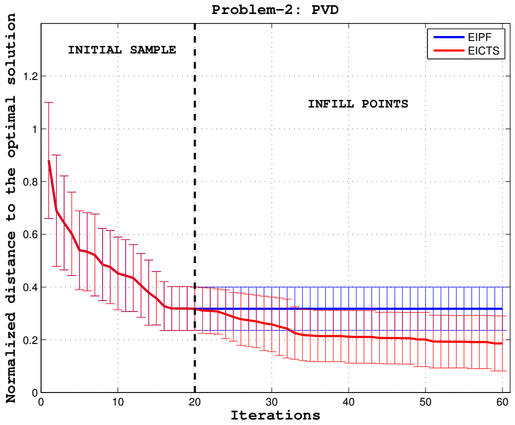

Fig. 5. Average of the smallest Euclidean distance to the optimal solution in 30 experiments for the PVD problem.

3) Welded Beam (WB): The last validation problem is taken from [36], [37], [38]. It has 4 design variables and 6 constraints. The performances in Fig. 6 look similar however the runtime performance of EICTS is approximately 3.5 times faster than EIPF. Moreover, Table IV shows the average number of feasible solutions found in 30 experiments for each problem and EIPF could not find any feasible solution, whereas EICTS could find at least 1 or 2 feasible solutions in approximately 10 out of 30 experiments.

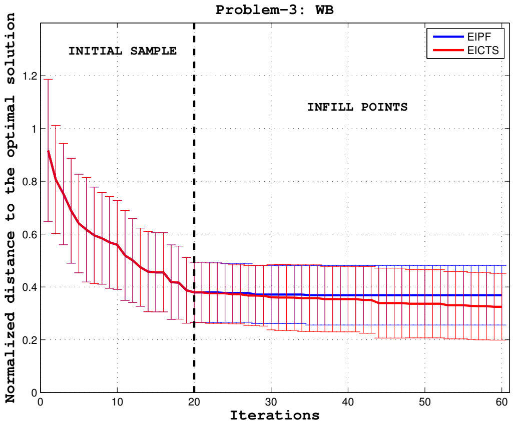

Fig. 6. Average of the smallest Euclidean distance to the optimal solution in 30 experiments for the WB problem.

B. Simulation-based Optimization Problem: FSW

Original problem was formulated as a multi-objective optimization problem in [15] and here it is converted to a single-objective optimization problem. The design variables are the radius of the tool shoulder ($8 \leq R_{sh} \leq 12$ mm), radius of the probe ($3 \leq R_{pr} \leq 6$ mm), welding speed ($1 \leq u_{weld} \leq 8$ mm/s) and the tool rotational speed in terms of revolutions per minute ($400 \leq n_{rev} \leq 1000$ rpm). As common in other manufacturing processes as well, the production rate is desired to be maximized in the FSW process; this can obviously be achieved by increasing the traversing speed (u_{weld}); however, increasing the welding speed promotes the risk of tool probe failure due to colder (i.e., harder) material ahead of tool. Therefore, temperature gradient in the narrow region of the tool probe surface (i.e., 1 mm ahead and one fourth of the height) should also be minimized for safety purposes. This evaluation can be done in more detail with a computational solid mechanics (CSM) or a computational fluid mechanics (CFD) approach; however, the computational cost would be very high (varies from hours to a day), and consequently, the integration of these type of models with the numerical optimization algorithms would be impractical. Hence, this safety issue has been considered as the objective function (i.e., to minimize), and thus bypassed the costly CSM/CFD calculations using a pseudo-mechanical link with an engineering intuition. Some of the other important constants in the FSW simulation are the thickness of the plate (7 mm), the height of the tool probe (6 mm), observed shear layer thickness at the shoulder and probe sides as well as at the tip of the probe are 2 mm, 0.5 mm and 0.25 mm, respectively. The welded plates are 100 mm-long and 50 mm-wide. Three constraints are defined: **1** and **2)** Average temperature in the

TABLE IV
PERFORMANCE COMPARISON OF EIPF AND EICTS, WITH RESPECT TO (1) AVERAGE NUMBER OF FEASIBLE SOLUTIONS, (2) AVERAGE COMPUTING TIME AND (3) BEST FEASIBLE FITNESS VALUE AMONG 30 EXPERIMENTS.

Performance Criterion	Problem	EIPF	EICTS
$n_{feas,avg}$	GTCD	41.467	20.300
	PVD	0.3667	1.2667
	WB	None	0.0333
	FSW	15.500	6.7500
$CPU - time_{avg}$	GTCD	225.95	214.63
	PVD	353.15	210.53
	WB	702.25	213.56
	FSW	5385.3	3797.8
$Fitness_{feas,best}$	GTCD	4,075,744.10	3,035,175.86
	PVD	6680.9	6049.2
	WB	None	5.195
	FSW	3.722	3.693

shoulder side shear layer should be between 450°C and 500°C respectively, **3)** the material temperature in front of the tool probe should be higher than 425°C.

The optimal solution to this FSW problem is [$R^*_{shoulder}$, R^*_{probe}, u^*_{weld}, n^*_{rev}] = [12 mm, 3 mm, 1 mm/s, 700 rpm]. This is also intuitive because promoting a large volume of heat generation by using a FSW tool having the maximum tool shoulder radius and minimum tool probe radius (i.e., Volume$_{Sh-side-SL} = \pi(R^2_{shoulder} - R^2_{probe})t_{Sh-side-SL}$) as well as keeping the traversing speed at its minimum value leaves only one process parameter to optimize, i.e., that is n_{rev}. Due to higher computational cost of the simulations, 10 experiments are performed instead of 30 only for this problem. The results are shown in Fig. 7 and it clearly shows the dominance of the EICTS results. EICTS clearly finds process and design parameters yielding lower temperature difference between the probe and the incoming material. Moreover, the variance in the average of best solutions in 10 experiments are much narrower which shows the reliability of the proposed method. The runtime performance (∼1.5 times faster) also supports the success of EICTS.

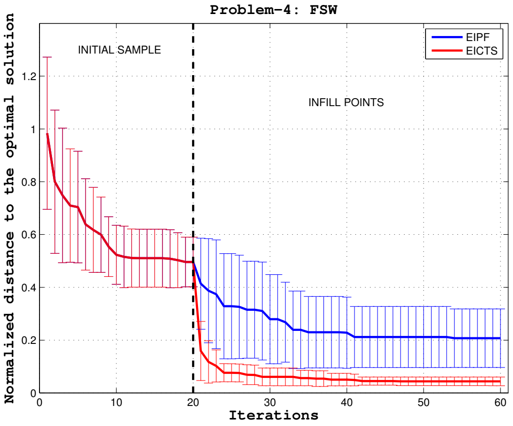

Fig. 7. Average of the smallest Euclidean distance to the optimal solution in 30 experiments for the FSW problem.

V. CONCLUSIONS

The steady-state thermal simulation of the Friction Stir Welding (FSW) process is implemented in COMSOL for optimizing the process parameters and the tool geometry to minimize the risk of tool failure (i.e., by minimizing the temperature gradient between the incoming material and tool probe). The average temperature is desired to be kept above 450°C to avoid tool pin failure and below 500°C to reduce the tool wear. The effect of the tool rotation on the distribution of the temperature field is also taken into account. Since the average computational time for a single FSW simulation is changing from 5 minutes to 9 minutes (on a desktop computer with Intel Core i7, 2.67 GHz CPU and 8 GB of RAM), these high fidelity simulations need to be executed sparingly. Therefore a surrogate-based optimization methodology, in which the computationally expensive function calls are replaced by low fidelity approximation functions (i.e., surrogates), is applied. A new constraint handling methodology, employing Constrained Tournament Selection, is developed and integrated within the well known efficient global optimization (EGO) framework. The runtime and convergence performance of the algorithm is tested on three analytical engineering test problems, and finally, on the simulation-based optimization problem in FSW. The performance of the proposed method, EICTS (Expected Improvement with Constrained Tournament Selection), is compared with EIPF (Expected Improvement with Probability of Feasibility). A budget of 60 high-fidelity simulations is allowed. Findings can be summarized as follows:

- Multiple number of constraint functions are approximated by only a single surrogate function, which is a different approach from most other surrogate-based optimization methods presented in the literature. This shortcut obviously speeds up the model building process.
- The order of magnitudes of any constraint can be handled with this method, because all constraint values are properly normalized without using any parameters (i.e., without knowing the maximum value of the constraint function can get, which is then used as a denominator constant). The proposed method does not need the actual constraint function violation values; instead it only compares the relative magnitudes of the total constraint violation (TCV) value within the Tournament Selection process to guide the search towards the feasible design region.
- EICTS has been tested on relatively difficult well known engineering test problems as well as the simulation based optimization problem for the FSW process. Both the runtime and convergence performances found to be very promising.

A few pointers to improve the performance of the proposed method could be given as following. Here, Kriging is used for both the objective and constraint functions, but the methodology can be extended to other surrogates (Support Vector Machines, Radial Basis Functions, etc.) as long as a measure of uncertainty (Mean Squared Error) estimation for the objective function is available. Such information can also be incorporated while handling constraints, i.e., instead of using only TCV_i, the value of $TCV_i + \hat{s_i}^2$ can be evaluated ($\hat{s_i}^2$ is the estimation of MSE for TCV_i). In that case, predictions at crowded regions would result in lower $\hat{s_i}^2$ values, whereas predictions at less explored regions would result in higher $\hat{s_i}^2$ values. Handling higher number of design variables is an important aspect of the constrained optimization practice, and it should be noted that the Bayesian Optimization (BO) Algorithms (such as EGO) are known to perform slower at higher number of dimensions (d $\geq$ 15). A *trust-region*-based BO methodology could be a good alternative for such a case. This will be addressed in a further study.

Acknowledgements

This research was supported in part by NSF under grant #DBI-0939454.

References

[1] Z. Ma, "Friction stir processing technology: a review," *Metallurgical and Materials Transactions A*, vol. 39, pp. 642–658, 2001.

[2] R. Mishra and Z. Ma, "Friction stir welding and processing," *Material Science and Engineering R*, vol. 50, pp. 1–78, 2005.

[3] Y. Chao and X. Qi, "Thermal and thermo-mechanical modeling of friction stir welding of aa6061-t6," *Journal of Materials Processing Manufacturing*, vol. 7, pp. 215–233, 1998.

[4] X. Zhu and Y. Chao, "Numerical simulation of transient temperature and residual stress in friction stir welding of 304l stainless steel," *Journal of Materials Processing Technology*, vol. 146, no. 2, pp. 263–272, 2004.

[5] H. Schmidt and J. Hattel, "Modelling heat flow around tool probe in friction stir welding," *Science and Technology of Welding and Joining*, vol. 10, pp. 176–186, 2005.

[6] ——, "Thermal modelling of friction stir welding," *Scripta Materialia*, vol. 58, pp. 332–337, 2008.

[7] P. Colegrove and H. Shercliff, "3-dimensional cfd modelling of flow around a threaded friction stir welding tool profile," *Journal of Materials Processing Technology*, vol. 169, no. 2, pp. 320–327, 2005.

[8] J. Robson, N. Kamp, and A. Sullivan, "Microstructural modelling for friction stir welding of aluminum alloys," *Materials and Manufacturing Processes*, vol. 22, no. 4, pp. 450–456, 2007.

[9] C. Chen and R. Kovacevic, "Parametric finite element analysis of stress evolution during friction stir welding," *Journal of Engineering Manufacture*, vol. 220, no. 8, pp. 1359–1371, 2006.

[10] D. Richards, Prangell, S. Williams, and P. Withers, "Global mechanical tensioning for the management of residual stresses in welds," *Materials Science and Engineering: A*, vol. 489, no. 1-2, pp. 351–362, 2008.

[11] Z. Feng, X. Wang, S. David, and P. Sklad, "Modelling of residual stresses in and property distributionsin friction stir welds aa6061-t6," *Science and Technology of Welding and Joining*, vol. 12, no. 4, pp. 348–356, 2007.

[12] C. Tutum and J. Hattel, *Multi-objective Evolutionary Optimisation for Product Design and Manufacturing*. Springer, 2011, ch. State-of-the-Art Multi-Objective Optimisation of Manufacturing Processes Based on Thermo-Mechanical Simulations, pp. 71–133.

[13] ——, "Numerical optimisation of friction stir welding: Review of future challenges," *Science and Technology of Welding and Joining*, vol. 16, no. 4, pp. 318–324, 2011.

[14] C. Tutum, K. Deb, and J. Hattel, "Hybrid search for faster production and safer process conditions in friction stir welding," in *8th International Conference on Simulated Evolution and Learning*, ser. SEAL'10, 2010, p. 603–612.

[15] ——, "Multi-criteria optimization in friction stir welding using a thermal model with prescribed material flow," *Materials and Manufacturing Processes*, vol. 28, no. 7, pp. 816–822, 2013.

[16] C. Tutum and J. Hattel, "Optimisation of process parameters in friction stir welding based on residual stress analysis: A feasibility study," *Science and Technology of Welding and Joining*, vol. 15, no. 5, pp. 369–377, 2010.

[17] R. Haftka, "Combining global and local approximations," *AIAA Journal*, vol. 29, no. 9, p. 1523–1525, 1991.

[18] Y. Jin, "A comprehensive survey of fitness approximation in evolutionary computation," *Journal of Soft Computing*, vol. 9, no. 1, pp. 3–12, 2005.

[19] N. Queipo, R. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, and P. Tucker, "Surrogate-based analysis and optimization," *Progress in Aerospace Sciences*, vol. 41, pp. 1–28, 2005.

[20] Y. Jin, "Surrogate-assisted evolutionary computation: Recent advances and future challenges," *Swarm and Evolutionary Computation*, vol. 1, no. 2, pp. 61–70, 2011.

[21] A. Forrester and A. Keane, "Recent advances in surrogate-based optimization," *Progress in Aerospace Sciences*, vol. 45, no. 1-3, pp. 50–79, 2009.

[22] J. Parr, A. Forrester, A. Keane, and C. Holden, "Enhancing infill sampling criteria for surrogate-based constrained optimization," *Journal of Computational Methods in Sciences and Engineering*, vol. 12, no. 1-2, pp. 25–45, 2012.

[23] R. Regis, "Constrained optimization by radial basis function interpolation for high-dimensional expensive black-box problems with infeasible initial points," *Engineering Optimization*, vol. 46, no. 2, pp. 218–243, 2014.

[24] P. Koch, S. Bagheri, W. Konen, C. Foussette, P. Krause, and T. Bäck, "A new repair method for constrained optimization," in *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '15, 2015, pp. 273–280.

[25] "Comsol," http://www.comsol.com, accessed: 2016-01-31.

[26] J. Davis, *ASM Specialty Handbook: Aluminum and Aluminum Alloys*. ASM International, 1993.

[27] R. Nandan and T. DebRoy, "Recent advances in friction stir welding process, weldment structure and properties," *Progress in Materials Science*, vol. 53, pp. 980–1023, 2008.

[28] D. Krige, "A statistical approach to some basic mine valuation problems on the witwatersrand," *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, vol. 52, no. 6, pp. 119–139, 1951.

[29] J. Sacks, W. Welch, T. Mitchell, and H. Wynn, "Design and analysis of computer experiments," *Statistical Science*, vol. 4, pp. 409–423, 1989.

[30] D. Jones, M. Schonlau, and W. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global Optimization*, vol. 13, p. 455–492, 1998.

[31] M. Sasena, "Flexibility and efficiency enhancements for constrained global design optimization with kriging approximations," Ph.D. dissertation, University of Michigan, 2002.

[32] F. Viana, R. Haftka, and V. Steffen, "Multiple surrogates: How cross-validation errors can helps us to obtain the best predictor," *Structural and Multidisciplinary Optimization*, vol. 39, no. 4, p. 439–457, 2009.

[33] F. Viana, "Multiple surrogates for prediction and optimization," Ph.D. dissertation, University of Florida, 2011.

[34] M. Schonlau, "Computer experiments and global optimization," Ph.D. dissertation, University of Waterloo, 1997.

[35] C. Beightler and D. Phillips, *Applied Geometric Programming*. New York: Wiley, 1976.

[36] E. Mezura-Montes and O. Cetina-Domínguez, "Empirical analysis of a modified artificial bee colony for constrained numerical optimization," *Applied Mathematics and Computation*, vol. 218, no. 22, p. 10943–10973, 2012.

[37] A. Hedar and M. Fukushima, "Derivative-free filter simulated annealing method for constrained continuous global optimization," *Journal of Global Optimization*, vol. 35, no. 4, pp. 521–549, 2006.

[38] K. Deb, "An efficient constraint handling method for genetic algorithms," *Computer Methods in Applied Mechanics and Engineering*, vol. 186, no. 2-4, pp. 311–338, 2000.