Antonis Kokossis, Michael C. Georgiadis, Efstratios N. Pistikopoulos (Eds.)
PROCEEDINGS OF THE 33rd European Symposium on Computer Aided Process Engineering (ESCAPE33), June 18-21, 2023, Athens, Greece
© 2023 Elsevier B.V. All rights reserved. http://dx.doi.org/10.1016/B978-0-443-15274-0.50209-2

Hybrid Modeling and Multi-Fidelity Approaches for Data-Driven Branch-and-Bound Optimization

Suryateja Ravutla[a], Jianyuan Zhai[b], Fani Boukouvala[a]

[a]*Department of Chemical and BiomolecularEngineering, Georgia Institute of Technology, Atlanta, GA 30332 USA*
[b]*Engineering & Data Sciences, Cargill Inc, Shanghai, 200031, China*
Email: sravutla3@gatech.edu

Abstract

High-Fidelity (HF) simulations are essential in quantitative analysis and decision making in engineering. In cases where explicit equations and/or derivatives are unavailable, or in the form of intractable nonlinear formulations, simulation-based optimization methods are used. We recently proposed a data-driven equivalent of spatial branch-and-bound that constructs underestimators of high-fidelity simulation data. Within this framework, low-fidelity surrogate data can also be used to inform underestimators. In this work, we utilize the recent advances in hybrid multifidelity surrogate modeling techniques to improve the validity of our underestimators, which leads to better bounds and incumbent optima with lower sampling requirements. Specifically, we show that by modeling the error between the high-fidelity and low-fidelity data, the surrogates learn more about the underlying function with less sampling requirements.

Keywords: Data-driven Optimization, Branch-and-bound, Convex underestimators, Multifidelity surrogate models, neural networks, support vector regression, hybrid modeling.

1. Introduction

In many engineering fields, it is desirable to simulate complex processes and use these quantifications of system performance for decision-making. Many such cases exist for chemical engineers, including molecular simulations, flowsheet simulations, computational fluid dynamic models, agent-based models, and more. These models are referred to as High-Fidelity (HF) simulations. But often these HF function evaluations are computationally expensive, and this implies that the number of function evaluations is limited by cost or time. Adding to this difficulty, in most cases, the objective functions and the constraints are only available as black-box evaluation function outputs (Fisher, Watson et al. 2020, van de Berg, Savage et al. 2022, Zhai and Boukouvala 2022). As a result, optimization of such systems becomes increasingly prohibitive. The absence of information on the system calls in for derivative-free optimization (DFO) or simulation-based optimization techniques. DFO techniques can be broadly classified into Sampling-based and Model-based methods. Numerous algorithms have been proposed in the recent years (Rios and Sahinidis 2013) for both cases. Sampling-based methods rely on comparing the function values directly and utilize this information to further sample adaptively. On the other hand, Model-based methods rely on constructing Machine Learning (ML) surrogate models as approximations of HF simulations with the aim to expedite the optimization process (Kim and Boukouvala 2020, Li, Dong et al. 2021). A disadvantage of using sampling-based methods is that they require too many samples to infer the optimum solution, while in surrogate model-based approaches, constructing an accurate surrogate can be challenging and computationally expensive. A solution that we have recently proposed combines advantages of both the methods in the form of a data-driven equivalent of the spatial branch-and-bound algorithm (DDSBB) (Zhai and Boukouvala 2022). The key idea of DDSBB is based on constructing convex underestimators of simulated data. A schematic of spatial branch-and-bound and its data-driven equivalent are shown is Figure 1.

These underestimators serve as relaxations and are convex, so they can be efficiently optimized, circumventing the task of directly optimizing nonconvex fitted surrogates. Samples drawn from the HF simulation serve as upper bounds (UB) of the global optimum and the minimum of the convex underestimator serve as lower bounds (LB). The search space is then progressively partitioned by using branching, node selection and pruning rules and adaptively sampling in the non-pruned subspaces. To build underestimators, using N samples, the underlying formulation is shown in Equation 1.

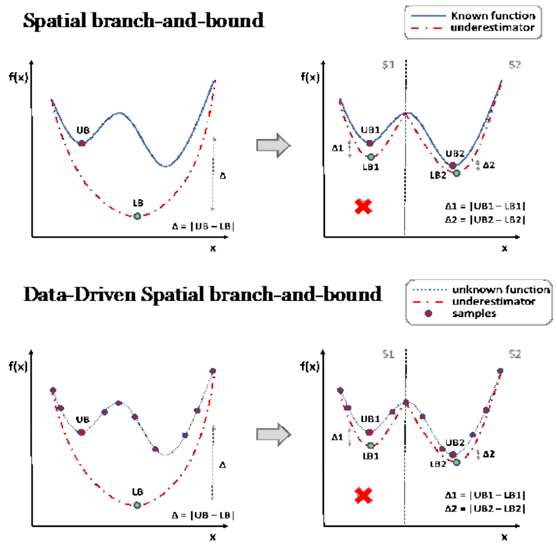


Figure 1: Deterministic Spatial Branch-and-Bound, and Data-Driven Spatial Branch-and-Bound, and the process of branching and bounding in both variants

$$\begin{aligned} &min_{a,b,c} \sum_i^N \left(f(\boldsymbol{x_i}) - f_{lb}(\boldsymbol{x_i})\right) \\ &s.t.\, f(\boldsymbol{x_i}) - f_{lb}(\boldsymbol{x_i}) \geq 0 \quad \forall\, i = 1 \; to\, N \\ &f_{lb}(\boldsymbol{x_i}) = \boldsymbol{a}\boldsymbol{x_i}^2 + \boldsymbol{b}\boldsymbol{x_i} + c \quad \forall\, i = 1 \; to\, N \\ &\boldsymbol{a} \geq 0, \qquad \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^D, \; c \in \mathbb{R} \end{aligned} \tag{1}$$

In contrast to the conventional sample-based and surrogate-based approaches, DDSBB employs some of the positive aspects of both. Like the sample-based approaches, it adaptively samples in the search space and uses this information to improve the bounding of the original function, by constructing the relaxations and pruning the subspaces that are not promising. It also utilizes surrogates to construct Low fidelity (LF) data but does not directly optimize the surrogates or rely on a single surrogate prediction. This LF data can be utilized along with the HF data to build the convex relaxations. In our recent work (Zhai and Boukouvala 2022), we have shown that by jointly using LF and HF data (multifidelity MF), we can optimize a higher fraction of benchmark problems

While the MF approach employed in DDSBB showed promising performance, a fraction of benchmark studies was still not optimized given limits on sampling requirements. In this current work, we improve the performance of our framework by incorporating more advanced hybrid multifidelity modeling techniques. Specifically, using the same amount of HF data as before, we attempt to learn more about our underlying black-box problem by modeling the error between the HF and LF data. The hypothesis is that this will overall improve our underestimators, and consequently the overall efficiency of the DDSBB approach, without increasing sampling requirements. Recent studies show that there is an increasing amount of focus in constructing surrogate models that combine data with different fidelity (Meng and Karniadakis 2020, Bradley, Kim et al. 2022). These Multifidelity Surrogate Models (MFSMs) exploit the relation between LF data and HF data. In this study we integrate these MFSMs into the DDSBB architecture and quantify their performance by benchmarking the method on 2-10 dimensional black-box optimization problems.

2. Overview and explored methods

2.1.1. ***Multifidelity approach in DDSBB***

The overview of DDSBB is shown in Figure 2. Initially, an input design is generated in the variable search domain by employing Latin Hypercube Sampling (LHS) to generate HF samples. This HF samples are then used for constructing quadratic underestimators by employing the formulation shown in Equation (1). This is shown in the figure using a solid red line. Subsequently, a set of branching, node selection and pruning rules are used to adaptively add samples in the non-pruned subspaces until convergence. Alternatively, as shown in red dotted line, if the MF approach is selected, these HF samples are used to generate LF surrogate models and LF samples. These LF samples are then used along with the HF samples to construct the underestimators. Currently, DDSBB has the capability to use Support Vector Regression (SVR), Neural Networks (NN) and Gaussian Process Regression (GPR) as the surrogate options. For the rest of the study, we utilize SVR as the surrogate option.

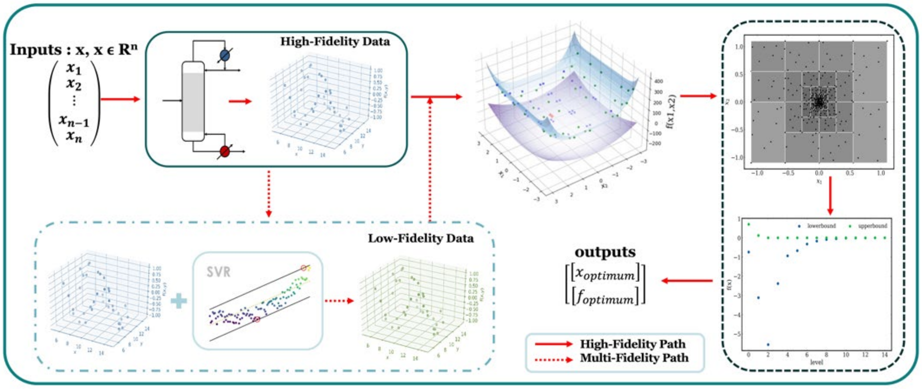


Figure 2: Overview of DDSBB. High-fidelity approach is shown in solid red line and multi-fidelity approach is shown in dotted red line.

2.1.2. ***Multifidelity Surrogate models to improve LF data***

There is an inherent trade-off between the number of HF samples that are used to construct LF surrogates, and the accuracy of the constructed surrogates. Also, adding the LF samples makes the convex underestimators more conservative with respect to bounding the function. We have found that in certain cases this leads to improving their validity, and as a result it leads to locating the global optimum of challenging benchmarks (Zhai and Boukouvala 2022). At the same time, LF surrogate predictions can make the underestimators overly conservative, thus leading to large sampling requirements for convergence. In this work, our hypothesis is that, MFSMs can be used to improve the accuracy of the LF data, which leads to improvement in underestimator validity. A widely used correlation to build MFSMs is: $y_H = \rho(x)y_L + \delta(x)$ where y_L, y_H represent the low and high-fidelity data respectively, $\rho(x)$ is multiplicative correlation surrogate and $\delta(x)$ is the additive surrogate. In a more general way, we can re-write it as $y_H = F(x, y_L)$. To establish a correlation between the HF and LF data, one would need a HF and a LF model to generate that data (Meng and Karniadakis 2020, Bradley, Kim et al. 2022). In case an LF model is available, it can be directly used. In cases where the LF model is not available, we propose a framework (Workflow 1) shown below to create a LF model using a fraction of available data.

Workflow 1: Constructing MFSMs

```
Let the data set $[\boldsymbol{x}_{HF}, y_{HF}]_{tot}$ represent the complete HF data.
Generate a training set data $[\boldsymbol{x}_{HF}, y_{HF}] = 75\%[\boldsymbol{x}_{HF}, y_{HF}]_{tot}$
TRAIN SVR
            Set $\boldsymbol{x}_{HF} \leftarrow$ input and $y_{HF} \leftarrow$ output,
            While termination criteria not true:
                Calculate $MSE_{SVR} = \frac{1}{N_{LF}}\sum_{i=1}^{N_{LF}}(|y_{LF} - y_{HF}|^2) + \beta_1 \parallel \phi_{SVR} \parallel_2$
                Tune SVR parameters
Generate LF dataset $[\boldsymbol{x}_{HF}, y_{LF}]_{tot}$ using HF input $[\boldsymbol{x}_{HF}]_{tot}$
Utilize the correlation $y_H = F(x, y_L)$ to model error between LF and HF outputs
$[y_{LF}]_{tot}$ and $[y_{HF}]_{tot}$ respectively using a NN
TRAIN NN
            Set $[\boldsymbol{x}_{HF}, y_{LF}]_{tot} \leftarrow$ input and $y_{HF} \leftarrow$ output,
            While termination criteria not true:
                Calculate $MSE_{NN} = \frac{1}{N_{HF}}\sum_{i=1}^{N_{HF}}(|y_{LF}^{*} - y_{HF}|^2) + \beta_2 \parallel \phi_{NN} \parallel_2$
                Tune NN parameters
```

In the Workflow 1, y_{LF}, y_{LF}^{*} represent output from SVR and NN. N_{LF}, N_{HF} is the number of training data points and total HF data points. β_1, β_2 represent the regularization weights and ϕ_{SVR}, ϕ_{NN} represent the associated parameters with SVR and NN respectively. A schematic for the Workflow 1 is shown in Figure 3.

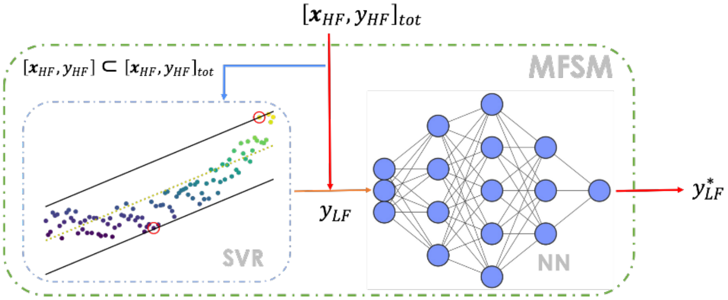

$[\boldsymbol{x}_{HF}, y_{HF}]_{tot}$
$[\boldsymbol{x}_{HF}, y_{HF}] \subset [\boldsymbol{x}_{HF}, y_{HF}]_{tot}$
MFSM
SVR
y_{LF}
NN
y_{LF}^{*}

Figure 3: A schematic of the workflow for fitting the MFSMs. Part of the total HF data set (shown in blue) is used to train LF SVR model and generate LF data y_{LF}

3. Results and discussion

To understand and visualize the effect of MF and MFSM approach on underestimator construction, we first take a 1- dimensional case study. Let us consider that the function $f(x) = sin(x) + sin(10x/3)$ is available as a black-box function, with $x \in [0, 9]$ and known global solution at $xopt = 5.14574$. Figure 4 shows underestimators constructed using HF, MF and MFSM approaches. We can see among all three approaches, the optimum solution identified is more accurate in MFSM approach and LF data in case of MFSM approach is more accurate in comparison to MF approach, without increasing the number of HF data collected.

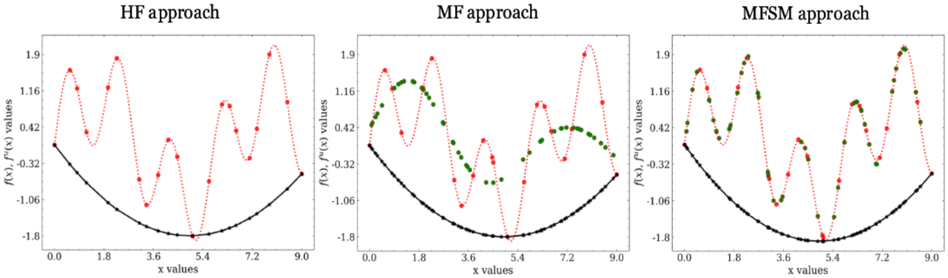
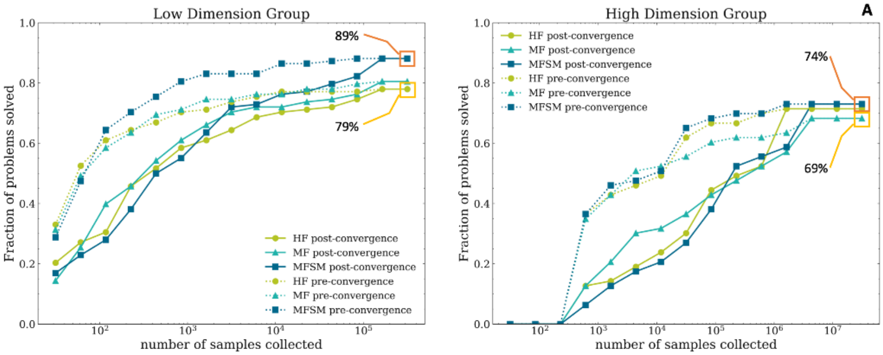


Figure 4: Comparison of constructed underestimators and LF data using HF, MF and MFSM approaches. Black-box function is shown in red dotted line, HF samples in red circles and LF samples in green circles. Underestimator is shown in solid black line.

Next, we test the performance of all three approaches on a large set of continuous box-constrained benchmark problems with known global solutions (BARON 2022). The benchmark problems were divided into two groups based on their dimensionality: lower dimensional group containing 118 problems with 2-3 variables and higher dimensional group containing 63 problems with 4-10 variables. All three approaches were initialized with 10*dimension+1 initial samples and converge when the $|UB - LB| \leq 0.05$ or $|UB - LB|/|LB| \leq 0.001$. For the performance analysis, we use the criterion $f_{best} \leq max(f^* + 0.01,\ (1.01)f^*)$ to allow a tolerance limit towards the optimal solution found. $f^*$and f_{best} represent the known global solution and the solution reported by DDSBB as optimum. We study the fractions of problems solved with sampling and CPU requirements based on pre and post convergence solutions. *Pre-convergence* and *post-convergence* solutions represent the optimum solution found by the algorithm *before* and *after* closing the UB-LB gap respectively. The performance curves for all three approaches are shown in Figure 5. In Figure 5A, we can see that the MFSM approach solves higher fraction of benchmark problems, and the pre-convergence solution quality is better in both low and high dimensions groups. In Figure 5B, we show the CPU requirements for the three approaches. The HF approach does not involve any surrogate fitting, so the CPU time largely corresponds to HF sampling. On the other hand, CPU time in MF and MFSM approaches also include surrogate modeling costs. As expected MFSM approach takes higher CPU time for fitting the complex MFSM structure. Thus, the advantage of this method is expected to be even greater when sampling cost increases.



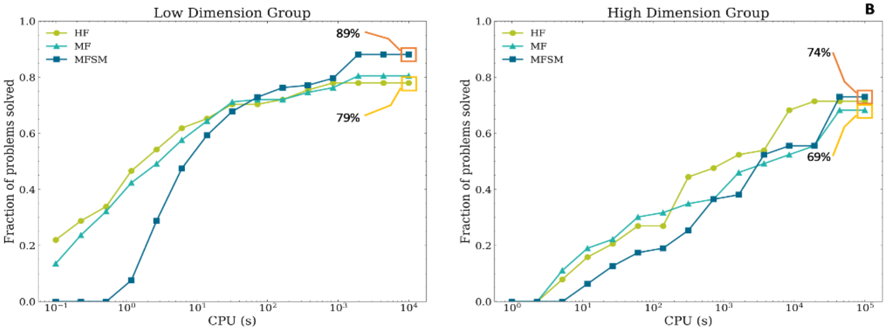


Figure 5: Performance curves of HF, MF, MFSM approaches. A) Fraction of problems solved vs sampling requirement, compared with pre and post convergence solutions reported by the algorithm. B) Fraction of problems solved vs CPU requirement.

4. Conclusions

In this work, we utilized multifidelity surrogate models as composite structures to model the error between HF and LF data to improve the validity of constructed data-driven underestimators embedded within a branch-and-bound framework. Results show that using composite/hybrid multifidelity models for surrogate-based optimization is promising, because it leads to more accurate surrogates with the same sampling cost but requires additional CPU time for training.

5. Acknowledgements

The authors acknowledge support from the National Science Foundation (NSF-1944678)

References

BARON (2022). Bound-constrained programs.

Bradley, W., J. Kim, Z. Kilwein, L. Blakely, M. Eydenberg, J. Jalvin, C. Laird and F. Boukouvala (2022). "Perspectives on the integration between first-principles and data-driven modeling." Computers & Chemical Engineering **166**: 107898.

Fisher, O. J., N. J. Watson, J. E. Escrig, R. Witt, L. Porcu, D. Bacon, M. Rigley and R. L. Gomes (2020). "Considerations, challenges and opportunities when developing data-driven models for process manufacturing systems." Computers & Chemical Engineering **140**: 106881-106881.

Kim, S. H. and F. Boukouvala (2020). "Surrogate-based optimization for mixed-integer nonlinear problems." Computers & Chemical Engineering **140**: 106847-106847.

Li, Z., Z. Dong, Z. Liang and Z. Ding (2021). "Surrogate-based distributed optimisation for expensive black-box functions." Automatica **125**: 109407.

Meng, X. and G. E. Karniadakis (2020). "A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse PDE problems." Journal of Computational Physics **401**: 109020-109020.

Rios, L. M. and N. V. Sahinidis (2013). "Derivative-free optimization: a review of algorithms and comparison of software implementations." Journal of Global Optimization **56**(3): 1247-1293.

van de Berg, D., T. Savage, P. Petsagkourakis, D. Zhang, N. Shah and E. A. del Rio-Chanona (2022). "Data-driven optimization for process systems engineering applications." Chemical Engineering Science **248**: 117135-117135.

Zhai, J. and F. Boukouvala (2022). "Data-driven spatial branch-and-bound algorithms for box-constrained simulation-based optimization." Journal of Global Optimization **82**(1): 21-50.